UNIVERSIDAD DE CÓRDOBA

Departamento de Informática y Análisis Numérico

Programa de doctorado en computación avanzada, energía y plasmas

Diagnóstico automático de melanoma mediante técnicas modernas de aprendizaje automático

Automatic melanoma diagnosis via modern machine learning techniques

Memoria de Tesis presentada por

Eduardo Pérez Perdomo

como requisito para optar al grado

de Doctor en Informática

Director

Dr. Sebastián Ventura Soto

Córdoba, Octubre de 2022

TITULO: *Automatic melanoma diagnosis via modern machine learning techniques*

AUTOR: *Eduardo Pérez Perdomo*

© Edita: UCOPress. 2022
Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

https://www.uco.es/ucopress/index.php/es/
ucopress@uco.es

UNIVERSITY OF CÓRDOBA

Department of Computer Science and Numerical Analysis

Automatic melanoma diagnosis via modern machine learning techniques

A Thesis submitted by

Eduardo Pérez Perdomo

in fulfilment of the requirements for
the degree of Doctor in Computer Science

Supervisor

Dr. Sebastián Ventura Soto

Córdoba, October 2022

La memoria titulada *"Diagnóstico automático de melanoma mediante técnicas modernas de aprendizaje automático"*, que presenta Eduardo Pérez Perdomo para optar al grado de Doctor en el marco del programa de doctorado "Computación avanzada, energía y plasmas", recopila un trabajo original de investigación realizado en el Departamento de Informática y Análisis Numérico de la Escuela Politécnica Superior de la Universidad de Córdoba, en el Instituto Maimónides de Investigación Biomédica de Córdoba, y la empresa Magtel. Dicho trabajo ha sido realizado bajo la dirección del Dr. Sebastián Ventura Soto y cumpliendo, a su juicio, los requisitos exigidos a este tipo de trabajos y respetando los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Córdoba, Octubre de 2022

El candidato:

Fdo.: Eduardo Pérez Perdomo

El director:

Fdo.: Dr. Sebastián Ventura Soto

TÍTULO DE LA TESIS:
Diagnóstico automático de melanoma mediante técnicas modernas de aprendizaje automático.

DOCTORANDO/A:
Eduardo Pérez Perdomo.

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

El trabajo realizado por el doctorando durante todo el período de investigación ha sido satisfactorio, lo cual se puede constatar en los resultados obtenidos. En primer lugar, se realizó un amplio estudio experimental sobre redes convolucionales para el diagnóstico de melanoma. La revisión dio lugar a la redacción de un artículo publicado en la revista *Medical Image Analysis.* Además, se presentó un poster presentado en el VII Congreso Científico de Investigadores en Formación de la Universidad de Córdoba. Dichos estudios permitieron una mejor comprensión del estado del arte, así como dar recomendaciones a la hora de elegir cuál modelo aplicar en un conjunto de datos.

A continuación, se comenzó a trabajar en varias líneas asociadas al desarrollo de modelos propios. En primer lugar, se implementó un modelo que combina submodelos y los rasgos abstractos que producen. Para realizar dicha combinación se empleó un algoritmo evolutivo, que codificaba un ensemble completo. El modelo produjo resultados competitivos, siendo comparado con los algoritmos que integran el estado del arte, y fue publicado en la revista *Neural Computing and Applications*. Asimismo, el modelo fue implementado en una aplicación web con el objetivo de ser intuitivo y sencillo de probar.

Se realizó una revisión de un nuevo tipo de arquitectura denominada *Dynamic routing between capsules*. Este nuevo tipo de arquitectura guarda mejor la posición relativa de los elementos que forman una imagen, en comparación con las redes convolucionales tradicionales. Como fruto de la investigación se ha construido una arquitectura basada en ambos tipos de arquitecturas. Se llevó a cabo un amplio estudio experimental y se llegó a la conclusión que obtiene mejores resultados que varias técnicas del estado del arte. Se publicó un artículo en la revista *Cancers*.

Se desarrollaron otras dos propuestas que están actualmente en proceso de revisión en las revistas *Artificial Intelligence in Medicine* y *Knowledge-based systems*. Dichas propuestas radican en generar imágenes de melanoma ficticias de alta calidad para remediar la falta de datos, y la segunda propuesta consiste en ajustar activamente el entrenamiento de los modelos a medida que aprende empleando *Active Learning*.

Es de destacar también que el trabajo de investigación desarrollado ha sido aceptado para su presentación en varios congresos científicos de prestigio internacional, tales como la *IEEE International Conference on Omni-layer Intelligent systems*. Por otra parte, las líneas de investigación desarrolladas en esta memoria no están aún agotadas, existiendo algunas líneas de trabajo futuro que considero pueden también dar lugar a varias publicaciones científicas de calidad.

Por todo lo comentado, considero que el trabajo presentado por D. Eduardo Pérez Perdomo reúne los requisitos exigibles a una tesis doctoral, y recomiendo la presentación de la misma.

Córdoba, 4 de octubre de 2022

Firma del director

Firmado por VENTURA SOTO
SEBASTIAN EMILIO - ***1000** el
día 04/10/2022 con un certificado

Fdo.: Sebastián Ventura

Tesis Doctoral subvencionada por los Contratos i-PFIS: doctorados IIS-empresa en ciencias y tecnologías de la salud de la convocatoria 2017 de la Acción Estratégica en Salud 2013 - 2016 (referencia **IFI17/00015**), convocatoria publicada en el B.O.E. del 25 de abril de 2017 con código BDNS №343490, resuelta provisionalmente por el Instituto de Salud Carlos III, con fecha 29 de septiembre de 2017, y de forma definitiva el 4 de diciembre de 2017[1].

Asimismo, esta Tesis Doctoral ha sido parcialmente subvencionada por el Ministerio de Economía y Competitivad (proyecto **TIN2017-83445-P**), el Ministerio de Ciencia e Innovación (Grant No. **PID2020-115832GB-I00**), y fondos FEDER (UCO-FEDER 18 **REF.1263116 MOD.A**).

[1] https://sede.isciii.gob.es/anouncements_detail.jsp?pub=10612

Agradecimientos

Mi agradecimiento al Instituto de Salud Carlos III por su apoyo al permitir la realización de esta tesis; al Instituto Maimónides de Investigación Biomédica de Córdoba por acogerme en sus instalaciones; al grupo empresarial Magtel durante el período de estancia; y al departamento de Informática y Análisis Numérico y a la Escuela Politécnica Superior.

A mi director Sebastián, por darme la oportunidad de desarrollar la tesis en su grupo Extracción de Conocimiento y Sistemas Inteligentes (KDIS). Tu experiencia y apoyo han sido claves en su consecución.

A mis compañeros de laboratorio, con quienes he tenido la suerte de compartir en los últimos años, José M., Rafa, Aurora R., Aurora E., Antonio T., Antonio M., Alejandro, Álvaro. A Jose L. y a todo el grupo KDIS por su apoyo.

A los amigos que han estado en los momentos precisos, Abel, Roger y Oscar.

Por encima de todo, a Dios y a mi familia pequeña, por su apoyo incondicional.

Gracias.
Eduardo

Resumen

Las tasas de incidencia y mortalidad del cáncer de piel siguen siendo una gran preocupación en muchos países. Según las últimas estadísticas sobre el cáncer de piel por melanoma, sólo en Estados Unidos se esperan 7.650 muertes en 2022, lo que representa 800 y 470 muertes más que en 2020 y 2021, respectivamente. En 2022, el melanoma se sitúa como la quinta causa de nuevos casos de cáncer, con un total de 99.780 personas. Esta enfermedad se diagnostica principalmente con una inspección visual de la piel y luego, si quedan dudas, se realiza un análisis dermatoscópico. El desarrollo de herramientas de diagnóstico no invasivas eficaces para las fases tempranas de la enfermedad debería aumentar la calidad de vida y disminuir los recursos económicos necesarios. El diagnóstico precoz de las lesiones cutáneas sigue siendo una tarea difícil, incluso para los dermatólogos expertos, debido a la complejidad, la variabilidad, la dudosa sintomatología y las similitudes entre las distintas categorías de lesiones cutáneas.

Para lograr este objetivo, trabajos anteriores han demostrado que el diagnóstico precoz a partir de imágenes de la piel puede beneficiarse enormemente del uso de métodos computacionales. Varios estudios han aplicado métodos basados en el análisis manual de las imágenes dermatoscópicas e histológicas de alta calidad y, además, técnicas de aprendizaje automático, como *k-nearest neighbors*, *support vector machines* y *random forest*. Sin embargo, hay que tener en cuenta que, aunque la extracción manual de características incorpora una importante base de conocimientos al análisis, la calidad de los descriptores extraídos depende en gran medida de la experiencia de los expertos. La segmentación de la lesión es otro proceso que suele realizarse manualmente. Ambos procesos consumen mucho tiempo y son propensos a cometer errores. Además, la definición explícita de una característica intuitiva e interpretable es difícilmente alcanzable, ya que depende del espacio de intensidad de los píxeles y, por tanto, no son invariables respecto a las diferencias en las imágenes de entrada. Por otra parte, el uso de dispositivos móviles ha aumentado considerablemente, lo que ofrece una fuente de datos casi ilimitada.

En los últimos años, se ha prestado cada vez más atención al diseño de modelos de *Deep Learning* para el diagnóstico del melanoma, más concretamente a las *Convolutional Neural Networks*. Este tipo de modelos es capaz de extraer y aprender características de alto nivel a partir de imágenes en bruto y/u otros datos sin la intervención de expertos. Varios estudios han mostrado que los modelos de aprendizaje profundo pueden superar a los métodos basados en la extracción manual de rasgos, e incluso igualar el rendimiento predictivo de los dermatólogos. Sin embargo, sugerimos cautela y una mayor integración de las propuestas en la práctica médica diaria, con el fin de validar plenamente tales hallazgos. La asociación *International*

Skin Imaging Collaboration ha impulsado el desarrollo de métodos para la obtención de imágenes digitales de la piel. Cada año, desde 2016 hasta 2019, se ha organizado un reto y una conferencia, en los que han participado más de 185 equipos. Sin embargo, los modelos convolucionales presentan varios problemas para el diagnóstico de la piel. Estos modelos pueden ajustarse a una gran diversidad de puntos de datos no lineales, siendo propensos a sobreajustarse en conjuntos de datos con un pequeño número de ejemplos de entrenamiento por categoría y, por tanto, alcanzando una pobre capacidad de generalización. Por otra parte, este tipo de modelo es sensible a algunas características de los datos, como las grandes similitudes interclase y las varianzas intraclase, las variaciones en los puntos de vista, los cambios en las condiciones de iluminación, las oclusiones y el desorden del fondo, que pueden encontrarse sobre todo en las imágenes no dermatoscópicas. Estos problemas suponen un reto para el diagnóstico automático de la enfermedad en las primeras fases.

Como consecuencia de lo anterior, el objetivo de esta tesis doctoral es realizar aportaciones significativas al diagnóstico automático del melanoma. Las propuestas pretenden evitar el sobreajuste y mejorar la capacidad de generalización de los modelos profundos, así como conseguir un aprendizaje más estable y una mejor convergencia. Hay que tener en cuenta que la investigación en aprendizaje profundo suele requerir una potencia de procesamiento abrumadora para entrenar arquitecturas complejas. Por ejemplo, cuando se desarrolló la arquitectura NASNet, los investigadores utilizaron 500 NVidia P100 - cada unidad gráfica costó entre $5.899 a $7.374, lo que representa un total de $2.949.500,00 - $3.687.000,00. Lamentablemente, la mayoría de los grupos de investigación no tienen acceso a tales recursos, incluido el nuestro.

En esta tesis doctoral se ha explorado el uso de varias técnicas. En primer lugar, se llevó a cabo un amplio estudio experimental, que incluyó modelos y métodos de última generación para aumentar el rendimiento. Se aplicaron técnicas bien conocidas, como *data augmentation* y *transfer learning.* Se llevó a cabo *Data augmentation* para equilibrar el número de instancias por categoría y actuar como regularizador para evitar el sobreajuste en las redes neuronales. Por otro lado, *transfer learning* utiliza los pesos de un modelo preentrenado de otra tarea, como estado inicial para el aprendizaje de una tarea distinta de la anterior. Los resultados demuestran que el diagnóstico automático del melanoma es una tarea compleja. Sin embargo, diferentes técnicas son capaces de mitigar estos problemas en cierto grado. Por último, se dan sugerencias sobre cómo entrenar modelos convolucionales para el diagnóstico del melanoma y se presentan futuras líneas de investigación.

A continuación, se aborda el descubrimiento de arquitecturas basadas en *ensembles* mediante el uso de algoritmos genéticos. La propuesta es capaz de estabilizar el proceso de entrenamiento. Esto es posible gracias a la búsqueda de combinaciones subóptimas de las características abstractas de los miembros del *ensemble*, que se utilizan para entrenar un bloque convolucional. A continuación, se entrenan varios bloques de predicción al mismo tiempo, y el diagnóstico final se consigue combinando todas las predicciones individuales. Investigamos empíricamente las ventajas de la propuesta, que muestra una mejor convergencia, mitiga el sobreajuste del modelo y mejora la generalización. Además, el modelo propuesto está disponible en Internet y puede ser consultado por expertos.

La siguiente propuesta se centra en el diseño de una arquitectura avanzada capaz de fusionar los bloques convolucionales clásicos y un modelo novedoso conocido como *Dynamic Routing Bet- ween Capsules*. Este enfoque aborda las limitaciones de los bloques convolucionales utilizando un conjunto de neuronas en lugar de una neurona individual para representar los objetos. Cada cápsula aprende una descripción implícita de los objetos, como la posición, el tamaño, la textura, la deformación y la orientación. Además, se lleva a cabo un ajuste de los parámetros principales para garantizar un aprendizaje eficaz con pocos datos de entrenamiento. Se realizó un amplio estudio experimental en el que la fusión de ambos métodos superó a seis modelos del estado del arte.

Por otro lado, se propone un método robusto para el diagnóstico del melanoma, que se inspira en las conexiones residuales y en *Generative Adversarial Networks*. La arquitectura es capaz de producir imágenes sintéticas fotorrealistas de la piel de 512 × 512, incluso con pequeños conjuntos de datos de imágenes dermatoscópicas y no dermatoscópicas como dominios del problema. De este modo, se abordan la falta de datos, los problemas de desequilibrio y los problemas de sobreajuste. Varios modelos convolucionales se entrenan y evalúan ampliamente utilizando las imágenes sintéticas, ilustrando su eficacia en el diagnóstico del melanoma.

Finalmente, se propone un marco de trabajo inspirado en el aprendizaje activo. La estrategia de consulta basada en el entrenamiento por lotes permite un proceso de entrenamiento más rápido, al aprender sobre la complejidad de los datos de forma implícita. Dicha complejidad nos permite ajustar el proceso de entrenamiento después de cada iteración, lo que lleva al modelo a conseguir un mejor rendimiento en un menor número de iteraciones en comparación con un entrenamiento aleatorio. A continuación, se evalúa el método de entrenamiento analizando tanto el valor informativo de cada imagen como el rendimiento predictivo de los modelos. Se realiza un amplio estudio experimental, en el que los modelos entrenados con la propuesta obtienen resultados significativamente mejores que los modelos de referencia.

Los resultados sugieren que todavía hay espacio para mejorar el diagnóstico de las lesiones cutáneas. Sin embargo, los datos estructurados de laboratorio, los datos narrativos no estructurados y, en algunos casos, los datos de audio u observacionales, son señalados por los radiólogos como claves durante la interpretación de la predicción. Esto es particularmente cierto en el diagnóstico del melanoma, donde el contexto clínico es a menudo esencial. Por ejemplo, síntomas como el picor y la disponibilidad de varias imágenes de una misma lesión cutánea durante un período de tiempo prueban que la lesión está creciendo, son muy probables para sugerir un cáncer. El uso de diferentes tipos de datos de entrada podría ayudar a mejorar el rendimiento de los modelos de predicción aplicados a la medicina. En este sentido, se ha propuesto un primer algoritmo evolutivo destinado a explorar datos multimodales multiclase, que superó al modelo base de tipo convolucional de una sola entrada. Además, las características predictivas extraídas por las cápsulas primarias podrían utilizarse para entrenar otros modelos, como *Support Vector Machine*.

Abstract

The incidence and mortality rates of skin cancer remain a huge concern in many countries. According to the latest statistics about melanoma skin cancer, only in the Unites States, 7,650 deaths are expected in 2022, which represents 800 and 470 more deaths than 2020 and 2021, respectively. In 2022, melanoma is ranked as the fifth cause of new cases of cancer, with a total of 99,780 people. This illness is mainly diagnosed with a visual inspection of the skin, then, if doubts remain, a dermoscopic analysis is performed. The development of effective non-invasive diagnostic tools for the early stages of the illness should increase quality of life, and decrease the required economic resources. The early diagnosis of skin lesions remains a tough task even for expert dermatologists because of the complexity, variability, dubiousness of the symptoms, and similarities between the different categories among skin lesions.

To achieve this goal, previous works have shown that early diagnosis from skin images can benefit greatly from using computational methods. Several studies have applied handcrafted-based methods on high quality dermoscopic and histological images, and on top of that, machine learning techniques, such as the k-nearest neighbors approach, support vector machines and random forest. However, one must bear in mind that although the previous extraction of handcrafted features incorporates an important knowledge base into the analysis, the quality of the extracted descriptors relies heavily on the contribution of experts. Lesion segmentation is also performed manually. The above procedures have a common issue: they are time-consuming manual processes prone to errors. Furthermore, an explicit definition of an intuitive and interpretable feature is hardly achievable, since it depends on pixel intensity space and, therefore, they are not invariant regarding the differences in the input images. On the other hand, the use of mobile devices has sharply increased, which offers an almost unlimited source of data.

In the past few years, more and more attention has been paid to designing deep learning models for diagnosing melanoma, more specifically Convolutional Neural Networks. This type of model is able to extract and learn high-level features from raw images and/or other data without the intervention of experts. Several studies showed that deep learning models can overcome handcrafted-based methods, and even match the predictive performance of dermatologists. *The International Skin Imaging Collaboration* encourages the development of methods for digital skin imaging. Every year since 2016 to 2019, a challenge and a conference have been organized, in which more than 185 teams have participated. However, convolutional models present several issues for skin diagnosis. These models can fit on a wide diversity of non-linear data points, being prone to overfitting on datasets with small numbers of training examples per class and, therefore, attaining a poor generalization

capacity. On the other hand, this type of model is sensitive to some characteristics in data, such as large inter-class similarities and intra-class variances, variations in viewpoints, changes in lighting conditions, occlusions, and background clutter, which can be mostly found in non-dermoscopic images. These issues represent challenges for the application of automatic diagnosis techniques in the early phases of the illness.

As a consequence of the above, the aim of this Ph.D. thesis is to make significant contributions to the automatic diagnosis of melanoma. The proposals aim to avoid overfitting and improve the generalization capacity of deep models, as well as to achieve a more stable learning and better convergence. Bear in mind that research into deep learning commonly requires an overwhelming processing power in order to train complex architectures. For example, when developing NASNet architecture, researchers used 500 × NVidia P100s - each graphic unit cost from \$5,899 to \$7,374, which represents a total of \$2,949,500.00 - \$3,687,000.00. Unfortunately, the majority of research groups do not have access to such resources, including ours.

In this Ph.D. thesis, the use of several techniques has been explored. First, an extensive experimental study was carried out, which included state-of-the-art models and methods to further increase the performance. Well-known techniques were applied, such as data augmentation and transfer learning. Data augmentation is performed in order to balance out the number of instances per category and act as a regularizer in preventing overfitting in neural networks. On the other hand, transfer learning uses weights of a pre-trained model from another task, as the initial condition for the learning of the target network. Results demonstrate that the automatic diagnosis of melanoma is a complex task. However, different techniques are able to mitigate such issues in some degree. Finally, suggestions are given about how to train convolutional models for melanoma diagnosis and future interesting research lines were presented.

Next, the discovery of ensemble-based architectures is tackled by using genetic algorithms. The proposal is able to stabilize the training process. This is made possible by finding sub-optimal combinations of abstract features from the ensemble, which are used to train a convolutional block. Then, several predictive blocks are trained at the same time, and the final diagnosis is achieved by combining all individual predictions. We empirically investigate the benefits of the proposal, which shows better convergence, mitigates the overfitting of the model, and improves the generalization performance. On top of that, the proposed model is available online and can be consulted by experts.

The next proposal is focused on designing an advanced architecture capable of fusing classical convolutional blocks and a novel model known as Dynamic Routing Between Capsules. This approach addresses the limitations of convolutional blocks by using a set of neurons instead of an individual neuron in order to represent objects. An implicit description of the objects is learned by each capsule, such as position, size, texture, deformation, and orientation. In addition, a hyper-tuning of the main parameters is carried out in order to ensure effective learning under limited training data. An extensive experimental study was conducted where the fusion of both methods outperformed six state-of-the-art models.

On the other hand, a robust method for melanoma diagnosis, which is inspired on residual connections and Generative Adversarial Networks, is proposed. The architecture is able to produce plausible photorealistic synthetic 512 × 512 skin images, even with small dermoscopic and non-dermoscopic skin image datasets as problem domains. In this manner, the lack of data, the imbalance problems, and the overfitting issues are tackled. Finally, several convolutional modes are extensively trained and evaluated by using the synthetic images, illustrating its effectiveness in the diagnosis of melanoma.

In addition, a framework, which is inspired on Active Learning, is proposed. The batch-based query strategy setting proposed in this work enables a more faster training process by learning about the complexity of the data. Such complexities allow us to adjust the training process after each epoch, which leads the model to achieve better performance in a lower number of iterations compared to random mini-batch sampling. Then, the training method is assessed by analyzing both the informativeness value of each image and the predictive performance of the models. An extensive experimental study is conducted, where models trained with the proposal attain significantly better results than the baseline models.

The findings suggest that there is still space for improvement in the diagnosis of skin lesions. Structured laboratory data, unstructured narrative data, and in some cases, audio or observational data, are given by radiologists as key points during the interpretation of the prediction. This is particularly true in the diagnosis of melanoma, where substantial clinical context is often essential. For example, symptoms like itches and several shots of a skin lesion during a period of time proving that the lesion is growing, are very likely to suggest cancer. The use of different types of input data could help to improve the performance of medical predictive models. In this regard, a first evolutionary algorithm aimed at exploring multimodal multiclass data has been proposed, which surpassed a single-input model. Furthermore, the predictive features extracted by primary capsules could be used to train other models, such as Support Vector Machine.

Preface

The Spanish legislation for Ph.D. studies, RD 99/2011, published the 28^{th} of January of 2011 (BOE-A-2011-2541), grants each Spanish University competencies to establish the necessary supervision and evaluation procedures to guarantee the quality of Ph.D. theses. As unique requirement for the defence, this national regulation indicates that the manuscript should be accompanied by a document detailing the complementary learning activities carried out by the student.

Accordingly, the University of Córdoba has a specific regulation for Ph.D. studies, approved by its governing board the 21^{th} of December of 2011. This regulation establishes two different modalities to elaborate the manuscript that the student, under the supervision of one or more Ph.D. advisors, has to present at the end of his/her doctorate studies. This Ph.D. thesis follows the modality described in the article no. 24 of the aforementioned regulation, referred as *Ph.D. thesis as a compendium of publications*. According to that article, the Ph.D. thesis can be presented as a compendium of, at least, three research articles published (or accepted for publication) in research journals of high quality, i.e. appearing in the first three quartiles of the Journal Citation Reports (JCR). If such a requirement is fulfilled, the manuscript has to include: an introduction to justify the thematic cohesion of the Ph.D. thesis; the hypotheses and objectives to be achieved, and how they are associated to the publications; full copy of the publications, and conclusions.

Following these guidelines, this Ph.D. thesis is organised as described next. Firstly, an introductory part is divided into five chapters. More specifically, Chapter 1 presents the background and state of the art of the research areas in which this Ph.D. thesis is framed. Next, the motivation, objectives and hypotheses are detailed in Chapter 2. Chapter 3 explains the research methodology, while an overview of the obtained results is presented in Chapter 4. Lastly, Chapter 5 discusses conclusions and future work. The second part of the document is comprised of three chapters. Chapter 6 includes the three main publications derived from this Ph.D. thesis. Chapter 7 compiles other journal publications associated to this Ph.D. thesis. Finally, Chapter 8 provides the list of conference publications.

Contents

List of Acronyms

ABCDE Asymmetry, Border, Color, Diameter and Evolving

RNA-seq Ribonucleic acid sequencing

DL deep learning

CNN Convolutional Neural Network

EHR Electronic Health Records

GANs Generative Adversarial Network

PGGANs Progressive Growing Generative Adversarial Network

DEAP Distributed Evolutionary Algorithms in Python

RT Research Tasks

H Hypotheses

CAPSNET Dynamic Routing Between Capsules Architecture

SM Scientific Methods

MCC Matthews Correlation Coefficient

GPU Graphics Processing Unit

GA Genetic Algorithm

SGD Stochastic Gradient Descend

SHAP Shapley Additive Explanations

LIME Local Interpretable Model-agnostic Explanations

Part I

Ph.D. Dissertation

1

Introduction

This chapter presents the fundamentals and state of the art of the research areas in which this Ph.D. thesis is founded. More precisely, firstly how to improve the automatic diagnosis of melanoma using data science is described. Then, an introduction to descriptors-based techniques is presented, including an overview of the Asymmetry, Border, Color, Diameter and Evolving (ABCDE) method. Lastly, Convolutional Neural Network (CNN) models for the diagnosis of melanoma is explored in depth in order to analyze the current state of the field.

1.1. Automatic melanoma diagnosis

Diagnosing diseases is a major process for analyzing what explains the symptoms of a patient. Physicians often use a physical examination and diagnostic tests in order to assess the patients. Diagnosis is a really challenging task since many signs and symptoms may often be nonspecific. In this sense, differential diagnosis plays an important role, where several possible explanations are compared, involving the analysis of correlation among variables followed by the recognition and differentiation of patterns [26].

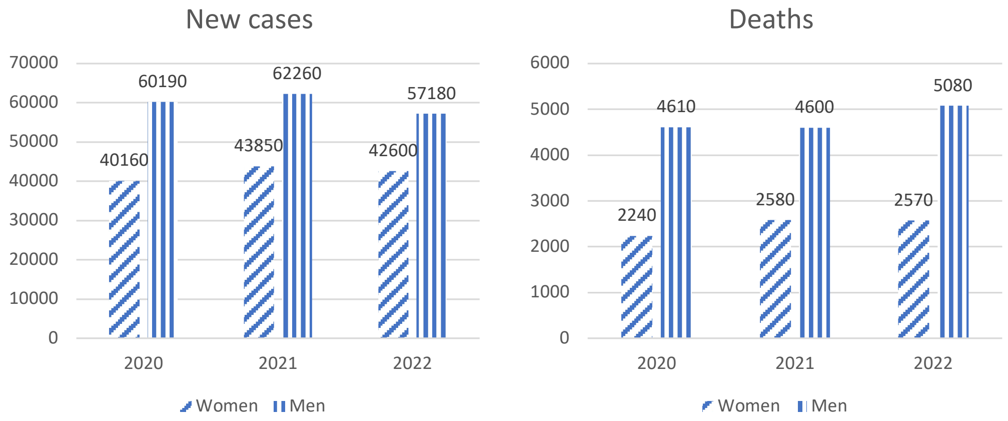


Figure 1.1: Statistics about skin cancer and prediction of the current year.

In recent years,the scientific community have been attracted by the application of artificial intelligence techniques [47, 49]. The above allows extracting key knowledge from hard scenarios where it could be impossible for humans to reach a remark. Loads of data is being stored, such as Electronic Health Records (EHR) and novel tests, which include massive gene array data. As a result, new applications are being developed for analyzing and predicting data in biomedicine [39].

We aims at researching about novel methods for diagnosing melanoma at early stages. Melanoma is leading to 90% of skin cancer mortality. Just in Europe, melanoma affects from 100 to 250 people per 1 million inhabitants; in USA from 200 to 300 per 1 million inhabitants; finally Australia from 500 to 600 per 1 million inhabitants [23, 25, 53]. According to the latest statistics (see Figure 1.1), this year in USA are due to occur 470 more deaths than 2021 [4–6].

The first step when diagnosing melanoma is to visually evaluate the lesion. After it, physicians conduct a deep analysis using a dermatoscope [3, 8]. However, the difference between malignant and benign lesions is sometimes too thin. As a consequence, a biopsy should be carried out. Nevertheless, there are methods that simplify such diagnosis. For example, the well-known as **A**symmetry, **B**order, **C**olor, **D**iameter, and **E**volution (ABCDE) method [1, 63]. Now, you can establish the stage and depth of the lesion (see Figure 1.2[1]).

[1] Taken from Cancer Research UK/Wikimedia Commons. https://bit.ly/3Cy70MF.

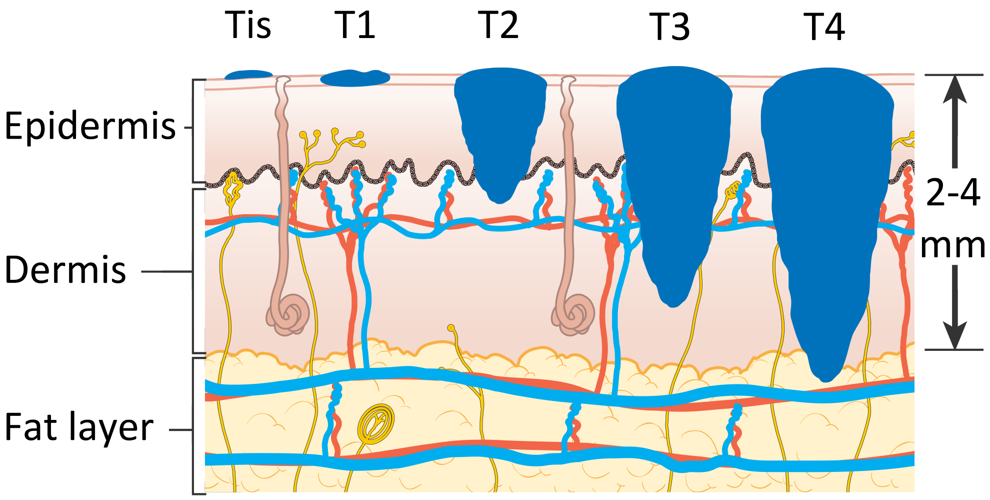
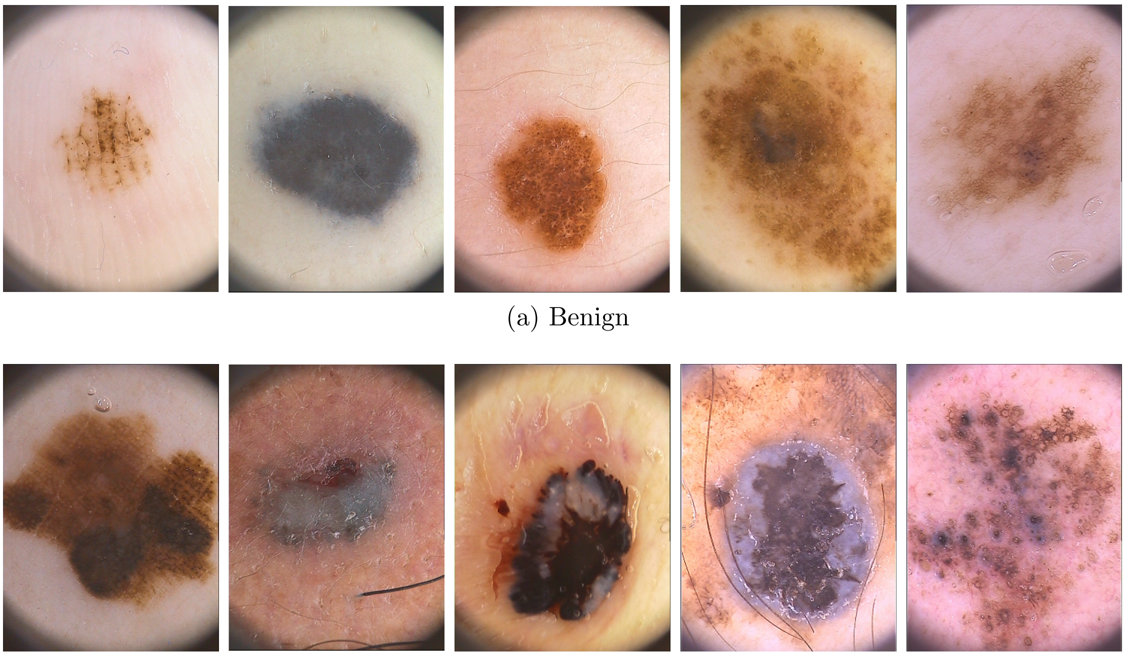


Figure 1.2: Stages of melanoma.

The early diagnosis of this illness is an arduous task even for a trained dermatologist or physician due to the variability and dubiousness of the symptoms [28], as can be seen in Figure 1.3. A bibliography review related to computational methods for melanoma diagnosis revealed that most state-of-the-art methods are based on classical machine learning techniques [7]. Most of these methods use as input information images of the nevi of each patient, or a set of descriptors extracted from those images, and apply classical techniques such as k-nearest neighbors [48, 55], artificial neural networks [11, 72], support vector machines [14, 27, 30, 86] and, more recently, deep learning models [21].

The classical techniques need the *priori* extraction of characteristics made by a dermatologist, which adds priceless insight about the problem. As a result, small datasets are required for building a machine learning model. Nevertheless, such effectiveness depends on the experts, and those set of high-level characteristics which are able to discriminate, are hard to find [42]. Such challenge arises because features are related to pixel intensity, and they are not invariant when analyzing differences between images [57]. As a consequence of the above, in recent years more and more attention has been paid to researching about novel artificial intelligence techniques which can identify and learn high-level features. In this sense, it could be tackled the inter- and intra-class variability in skin lesion images [2, 21].



(a) Benign

(b) Malignant

Figure 1.3: Morphological differences in images belonging to patients which have the same and different biological conditions; images taken from PH2 dataset [62].

All of these proposals use information of one or multiple nevi for each patient, or clinical information extracted from other sources. Furthermore, the number of proposals that make use of more flexible representations at the same time, such as clinical, gene and image data, is largely reduced [64]. For example, only a sample image is usually associated with one patient, which disables classical multi-instance learning. Also, only few dataset contains some clinical meta data, such as sex, lesion localization, and age. Even at this point, more than 11% of the samples contain missing metadata values, hampering the classical multi-view learning. In addition, after performing a deep analysis of the existing data sources, we have concluded that the majority of public skin cancer data sources only encompass a limit number of images, which reduces the capacity of machine learning algorithms.

In this sense, in recent years more care has been given to developing deep learning (DL) models for diagnosing melanoma, more specifically Convolutional Neural Networks [21], mainly because these models can automatically analyze and learn characteristics from skin lesions for a better prediction.

These facts motivate our work hypothesis, that is, that the use of novels techniques over these models will increase the overall diagnostic performance. The proposals will provide better machine learning models, which will be more accurate, robust, and probably more scalable and appropriate for use in a big data environment, e.g. by means of applying and developing new data augmentation methods [67], utilizing transfer learning techniques [21], designing CNN-based ensembles [60], developing methods and architectures to construct invariant models [56], designing stable training methods [68].

1.2. Convolutional Networks for the diagnosis of melanoma

CNNs are a type of feedforward neural network that use a mathematical operation called convolution, which is a linear operation that involves the multiplication of a two-dimensional input (an image) with two-dimensional array of weights, called a filter or a kernel. Convolution allows us to draw characteristics from skin lesions images automatically. The above are transferred to successive layers that can learn abstract characteristics, and at the end a final decision is made [51]. The above would mean that CNNs are able to construct complex concepts from a low level. For example, a human face could be interpreted as components (mouth and node), and the latest ones into sub-components (corners and contours) [33].

Several works have shown CNNs as a suitable technique for the automatic diagnosis of melanoma [21, J3], demonstrating that this type of deep learning model is more suitable than previous methods that need hand-crafted features. However, they still present various disadvantages that hamper their application for diagnosing melanoma. Firstly, CNNs can fit on a wide diversity of non-linear data points, and tend to overfitting on small data sources, and therefore, such models are very likely to attain poor predictive performance. In this regard, it is noteworthy that most of the majority of skin lesion data sources contains a limited number of images, reducing the abilities of CNNs. Secondly, CNNs are sensitive to large inter-class similarities and intra-class variances, variations in viewpoints, changes in lighting conditions, occlusions, and background clutter [9]. Although dermoscopic images are commonly used by expert dermatologists, nowadays there is an increasing tendency to get images shot with cheaper technological devices [20]. As a result, economical resources and invasive treatments can be reduced. In addition, new models should obtain suitable predictive performance in both non-dermoscopic and dermoscopic images.

Finally, CNNs are more tolerant to small changes in the viewpoint and roughly invariant to small translations on the training data. However, they are not rotation, color or lighting-invariant [54, 73]. Invariance to a transformation is key for the image recognition task, and this means that your input image could be transformed and the representation you get is the same as the original. The great variety of possible morphologies is a key task for improving the automatic diagnosis of melanoma. The model should be able to detect rotations, changes, and it is supposed to make internal modifications in the above sense.

On the other hand, in the past few years, several techniques have been applied to improve such predictive performance. For instance, to build a transformation-invariant model, data are augmented by applying custom transformations on the images [83]. For example, Lenc and Vedaldi [54] generated more transformation-invariant CNN models, where the benefits were more noticeable in deeper models. Also, it is well-known that combining a pretrained Inception model [77] and data augmentation can match and some times improve the performance of experts [21]. Recently, more advanced techniques can tackle the lack of data by augmenting it in a realistic manner [34], such as Generative Adversarial Network (GANs). GANs are able to augment a dataset by using two models. The generator creates data by randomly selecting points from the latent space. On the other hand, the discriminator determines whether a sample is a fake (sample created by the generator) or not (real sample from the dataset).

In addition, transfer learning is a method which uses knowledge from other tasks [58]. In the above scenarios are usually available loads of data. Transfer learning have been applied successfully for diagnosing melanoma. Romero *et al.* [70] proposed a VGG-based convolutional network to classify a lesion as malignant or benign. According to the authors, the sensitivity rate increased from 51% to 79% when applying transfer learning and fine-tuning part of the network. Mahbod *et al.* [61] introduced a new multi-scale multi-CNN fusion proposal based on ensemble learning. The authors applied and studied the impact of changing the resolution of an image for reusing knowledge. Results pointed out that the diagnosis performance could be improved by carefully selecting the image resolution.

On the other hand, ensemble learning [19] is used to solve complex tasks, since combining weak models from different areas are supposed to achieve better predictive performance than simpler models [38]. For instance, an ensemble-based approach, in which eight transformed images were generated from each original image and were then passed to two CNNs based on different architectures, was capable of obtaining competitive results with respect to the state-of-the-art methods for melanoma diagnosis [60]. Furthermore, an ensemble composed of members inspired on several CNN architectures, such as AlexNet [50], VGG and InceptionV1 [78], was applied on the International Symposium on Biomedical Imaging (ISBI) 2017 challenge and achieved suitable predictive performances [37].

Nevertheless, it is considered that the application of the above methods in the diagnosis of melanoma could be further improved.

2
Objectives

Dermatologists rarely achieve test sensitivities greater than 80%, despite their experience. This is because of the great variety of possible morphologies of moles, which is an important issue to solve in order to attain a more effective melanoma diagnosis. Several studies have shown that the early diagnosis of melanoma from skin images can benefit greatly from using computational methods [21].

As a consequence, we formulated the scientific problem: How do we increase the performance of predictive models which exploit the limited available information in skin lesion analysis, in order to obtain significantly better results than state-of-the-art algorithms?

The general objective of this thesis was to develop new algorithms and architectures with a high performance in the resolution of the automatic diagnosis of melanoma. In addition, the above approaches were made available through a web application.

This general objective was divided into the following sub-objectives:

- **O_1:** Analysis of the state-of-the-art algorithms in the diagnosis of melanoma, especially in the area of CNNs, to identify open problems and the techniques that could be applied to their resolution.

- **O_2:** Redesign of a segmentation method. It is well known that preprocessing the input data can improve the quality of biomedical data analyses. As a consequence, the CNN models can focus on the lesion.
- **O_3:** Design of genetic algorithms to support the discovery of novel CNN architectures.
- **O_4:** Design and development of an interactive web application in order to assist dermatologists in decision making.
- **O_5:** Design of advanced architectures, which are able to preserve hierarchical spatial relationships, such as position, size and texture.
- **O_6:** Design of an advanced data augmentation technique, which could not only improve performance of CNNs, but can also be used to train physicians.
- **O_7:** Design of a new training method in order to improve the effectiveness and stability of the models.

After a bibliographic review, the following Hypotheses (H) were formulated:

- **H_1:** If a deep analysis of the current state of the automatic diagnosis of melanoma was carried out, methods and algorithms could be employed as a baseline in order to achieve better diagnostic performance.
- **H_2:** If a preprocessing method removed irrelevant areas from the lesion, the performance of CNNs would improve.
- **H_3:** If a genetic algorithm designed an ensemble-based architecture for the diagnosis of melanoma, it would provide a good trade-off between efficiency and effectiveness in its solution.
- **H_4:** If a genetic algorithm discovered a multimodal architecture for the diagnosis of melanoma, it would add an extra-value that is commonly desired by experts.
- **H_5:** If a new architecture capable of preserving hierarchical spatial relationships were applied in melanoma diagnosis, it would obtain significantly better performance than state-of-the-art models.

- **H_6:** If a new data augmentation technique, which is able to generate plausible photorealistic synthetic 512 × 512 skin images, were applied to train CNN models, it would achieve significantly better performance than state-of-the-art data augmentation methods.
- **H_7:** If a framework, which is capable of combining several techniques, guided the training process, the performance of CNNs would improve.

In order to achieve the specific objectives and to test the hypotheses formulated, the following Research Tasks (RT) were accomplished:

- **RT_1:** Analyze the basis of the diagnosis of melanoma and review the state-of-the-art algorithms, identifying open problems.
- **RT_2:** Design and implement a new segmentation method on skin lesion images.
- **RT_3:** Validate the effectiveness of the proposed segmentation algorithm in the extraction of "relevant" pixels.
- **RT_4:** Validate the effectiveness of the proposed segmentation algorithm in the improvement of CNN.
- **RT_5:** Design and implement a genetic algorithm capable of building ensembles of CNN for melanoma diagnosis.
- **RT_6:** Validate the effectiveness of the proposed ensemble by means of comparing with the state-of-the-art CNN.
- **RT_7:** Design and implement an architecture capable of combining the best features from CNN and Dynamic Routing Between Capsules Architecture (CAPSNET) for melanoma diagnosis.
- **RT_8:** Validate the effectiveness of the proposed architecture by means of comparing with the baseline models and the state-of-the-art CNN.
- **RT_9:** Design and implement a genetic algorithm capable of discovering novel multimodal architectures for melanoma diagnosis.
- **RT_{10}:** Validate the effectiveness of the proposed multimodal model.

- **RT$_{11}$:** Design and implement an advanced data augmentation technique in order to cope with the imbalance and lack of data issues.

- **RT$_{12}$:** Validate the effectiveness of the proposed architecture by means of comparing with the state-of-the-art data augmentation techniques.

- **RT$_{13}$:** Design and implement a pipeline capable of training CNN models for melanoma diagnosis.

- **RT$_{14}$:** Validate the effectiveness of the proposed framework by means of comparing with the baseline state-of-the-art CNNs.

In the execution of these tasks, the following Scientific Methods (SM) were used:

- **SM$_1$:** General methods: the hypothetico-deductive method was used to elaborate the hypotheses and to propose research lines from partial results; the systematic method for the development of computational tools; the bibliographic revision method for the analysis of previous works.

- **SM$_2$:** Logic methods: the method of analysis and synthesis to decompose the information in logical and related parts, simplifying the information to process; the modeling method in the designing of algorithms and computational tools.

- **SM$_3$:** Empirical methods: the proposed algorithms and models were assessed by using experimentation.

- **SM$_4$:** Mathematical methods: statistical tests to validate the quality of the results.

3

Methodology

This chapter summarizes the methodology followed in this thesis, which mainly concerns the experimental evaluation of the different algorithms proposed. Specific information about the methodology employed in each experimental study is provided in its respective article.

Skin melanoma datasets

Table 3.1 shows a summary of the characteristics of the datasets used in the experimental study. The datasets were obtained from the following sources: *The International Skin Imaging Collaboration*[1] (ISIC) repository, PH2[2], MED-NODE[3], DERM7PT[4], SD-198[5] and The Dermofit Image Library[6] [10] (DERM-LIB). The ISIC repository includes the datasets HAM10000 [80], MSK [17], UDA [35], and BCN20000 [18]. PH2 comprises high-quality dermoscopic images, where manual

[1] https://www.isic-archive.com

[2] https://www.fc.up.pt/addi/ph2%20database.html

[3] http://www.cs.rug.nl/imaging/databases/melanoma_naevi/

[4] http://derm.cs.sfu.ca

[5] https://bit.ly/2TdZWQ6

[6] https://bit.ly/3gi6rKR

Table 3.1: Summary of the benchmark datasets.

Dataset	Source	# Img	ImbR	IntraC	InterC	DistR	Silho
BCN20000	[18]	17,393	2.848	9,014	10,107	0.892	0.153
DERM-LIB	[10]	407	4.355	7,171	9,163	0.783	0.270
DERM7PT-C	[46]	827	2.282	15,442	16,318	0.946	0.086
DERM7PT-D	[46]	827	2.282	15,971	16,866	0.947	0.087
HAM10000	[80]	7,818	6.024	8,705	9,770	0.891	0.213
ISBI2016	[35]	1,273	4.092	10,553	10,992	0.960	0.101
ISBI2017	[17]	2,745	4.259	9,280	9,674	0.959	0.089
MED-NODE	[31]	170	1.429	9,029	9,513	0.949	0.068
MSK-1	[17]	1,088	2.615	11,753	14,068	0.835	0.173
MSK-2	[17]	1,522	3.299	9,288	9,418	0.986	0.062
MSK-3	[17]	225	10.842	8,075	8,074	1.000	0.112
MSK-4	[17]	943	3.366	6,930	7,162	0.968	0.065
PH2	[62]	200	4.000	12,688	14,928	0.850	0.210
SDC-198	[76]	648	4.735	14,054	14840	0.947	0.116
UDA-1	[35]	557	2.503	11,730	12,243	0.958	0.083
UDA-2	[35]	60	1.609	11,297	11,601	0.974	0.020
Ave.		-	3.746	10,921	11,705	0.937	0.109

segmentation, clinical diagnosis and the identification of several dermoscopic structures were performed by expert dermatologists [62]. MED-NODE contains 170 non-dermoscopic images taken with mobile phones [31]. SDC-198 is a benchmark dataset for visual recognition of skin diseases that contains 6,584 real-world images from 198 categories [76]. DERM7PT-D and DERM7PT-C contain dermoscopic and clinical images, respectively [46]. Finally, DERM-LIB gathers 1,300 high-quality focal skin lesion images under standardized conditions. Each image has a diagnosis based on expert opinion and like PH2, includes a binary segmentation mask that denotes the lesion area.

All datasets were filtered in order to only analyze images belonging to the malignant and benign classes, e.g. BCN20000 and SDC-198. Finally, a total of 36,703 raw images were processed in the experimental study. ImbR means the imbalance ratio between the benign and malignant classes; IntraC and InterC represent the average distance between samples of the same class and the average distance between samples of different classes, respectively; DistR means the ratio between IntraC and InterC;

Silho indicates the silhouette score. It is noteworthy that the number of benign samples is several orders of magnitude higher than the number of malignant samples, which can hamper the learning process. Also, the average intra-class and inter-class distances were computed using the Euclidean distance function, where each image was represented as a vector of numbers. The ratio between these two distances indicated that both images of different classes, and images of the same class, are far from each other, so therefore denoting a high degree of overlapping between classes. Finally, the silhouette score [71] was also calculated, which represents how similar is an image to its own cluster compared to others. The results showed that images did not match well to their own cluster, and even samples belonging to different clusters were close in the feature space.

Software and hardware

The experimental studies were executed on several platforms:

- Ubuntu 18.04, 4 × Intel Core i7-8700K Processor 64 GB DDR4 RAM, 4 × GPUs Geforce GTX 1080-Ti, 4 × GPUs NVIDIA Geforce RTX 2080-Ti, 3 × GPUs NVIDIA Geforce GTX 780, 1 × GPUs NVIDIA Tesla K40c.
- Cluster of 12 compute nodes with Rocks cluster 6.1 x64 Linux distribution, Intel Xeon CPUs (E5645 and E52620) with 12 cores, 24 and 64 GB DDR memory.
- The experiments were implemented in Python v3.X, and the algorithms were developed by using Scikit Learn[65], Distributed Evolutionary Algorithms in Python (DEAP) [22], Tensorflow[7], MXNet[8] and PyTorch[9] frameworks.

Performance evaluation

Regarding the evaluation process, a 3-times 10-fold cross validation process was performed to assess the effectiveness of each model. In each fold, Matthews Correlation

[7] https://www.tensorflow.org/
[8] https://mxnet.apache.org/versions/1.8.0/
[9] https://pytorch.org/

Coefficient (MCC) was used to measure the predictive performance of the models; the higher the MCC value, the better the performance. MCC is widely used in Bioinformatics as performance metric [12, 16, J3], and it is specially designed to analyze the predictive performance on unbalanced data. The formulation of this measure, as also its interpretations, can be consulted in the articles derived from this thesis.

In addition, non-parametric statistical tests were used to detect whether there was any significant difference in predictive performance. Friedman's test [24] was conducted in cases where a multiple comparison was carried out, and afterward Hommel's post-hoc test [40] was employed to perform a multiple comparison with a control method. The Shaffer post-hoc test [74] was employed to perform pairwise comparisons. The Wilcoxon Signed-Rank test [84] was performed in those cases where only two individual methods were compared. All hypothesis testing was conducted at 95% confidence.

4
Results

This chapter presents the proposed models and summarizes the most relevant results. Firstly, an analysis is conducted, aimed at illustrating the most relevant state-of-the-art techniques for the automatic diagnosis of melanoma. An extensive experimental study was carried out in order to find how suitable CNNs are for the above task. Secondly, the genetic algorithm to address the discovery of CNN-based ensemble architectures is described. Lastly, the main features of the architecture created by fusing CAPSNET and CNN, which allows us to achieve a better diagnosis performance, are explained. The associated publications to each part are listed in the corresponding section, and the full content of the journal publications can be found in Part II of this document.

4.1. Experimental review

The main aim of this thesis is the development of accurate methods for the automatic diagnosis of melanoma, and in this chapter we provide a background in this problem. This information is useful as a starting point for developing new models. First, public available datasets were studied, which concluded that imagining data dominated, and only one Ribonucleic acid sequencing (RNA-seq) dataset was found, which comprises only tumor and metastasis samples. Nevertheless, a method has been developed as part of a collaboration with researchers from the Maimonides Biomedical Research Institute of Cordoba, which exploit the transcript-base data.

This method introduces a supervised machine learning-based methodology, allowing the determination of subsets of relevant gene/isoform components that best discriminate samples [J2]. The experimental results illustrate the utility and benefit of the proposed methodology for analyzing dysregulation in splicing machinery. Overall, experimental results indicate that the gene/isoforms expression-based rankings contain valuable information, which could lead to a more effective cancer diagnosis. However, due to the lack of data and the non-existence of the normal category, no further research was conducted on RNA-seq.

Nowadays, efforts are aimed at the early diagnosing of skin lesions by analyzing images, which vastly increase chances for cure and it is cheaper than invasive methods. Regarding the knowledge of dermatologists in this task, specialists achieve sensitivity levels of up to 80% using dermoscopic images [13], which increases the need to develop efficient methods that facilitate diagnosis and aid dermatologists in decision-making. In the literature, several studies have performed an experimental analysis on the techniques for the automatic diagnosis of melanoma, mainly focusing on descriptor-based methods [29] and CNNs models [J3]. Descriptor-based methods require the previous extraction of handcrafted features, which is time-consuming and relies on the expertise of dermatologists, introducing a margin of human error [42, 57]. On the other hand, CNNs models are able to learn a set of abstract features from raw images and achieve a high performance without the need for extracting handcrafted features [21, 87]. Nevertheless, in most cases, researchers trust their expertise to select which techniques to apply, do not follow a standard experimental study, and include a limited number of data. Therefore, given the absence of extensive experimental analysis comparing the performance of state-of-the-art CNNs, we carried out a study and provide suggestions on how to boost the diagnostic performance [J3].

While reviewing the state-of-the-art CNN models, we found several architectures and techniques which have previously been used in melanoma diagnosis. Firstly, the architectures are explained, aiming to shed light on what components help most for the diagnosis of melanoma. In addition, a ranking is shown by analyzing the number of trainable parameters. However, several issues were found when we tried to apply CNN models in the diagnosis of melanoma. For example, several of the existing public skin images datasets include only a few hundred images, e.g. PH2 and MSK-3 datasets with 200 and 225 images, respectively. Commonly, CNN models can fit on a wide diversity of non-linear data points, thus being prone to overfitting on such datasets with small numbers of training examples per class and, therefore, attaining a poor generalization capacity. In addition, although there is a growing tendency to collect images taken with common digital cameras, the number of non-dermoscopic images properly labeled is far too low compared to dermoscopic images [20]. On the other hand, CNNs are sensitive to large inter-class similarities and intra-class variances, variations in viewpoints, changes in lighting conditions, occlusions, and background clutter [9]. As a consequence, several techniques have been developed in order to improve the diagnostic performance.

Secondly, the most feasible and proven techniques for attaining a better melanoma diagnosis were analyzed, such as data augmentation [67], transfer learning techniques [21], CNN-based ensembles [60], or methods to construct invariant models [56].

Thirdly, an extensive experimental study was carried out, which compared several baseline CNNs to the same ones but evaluating different methods and techniques: a) optimization algorithms, b) weight balancing, c) transfer learning, d) data augmentation. Overall, a consensus was not found when analyzing the optimization algorithms and weight balancing techniques - specific behaviors were found in each CNN model. On the other hand, transfer learning and data augmentation proved to be suitable techniques for achieving significantly better diagnostic performance.

Fourthly, the complexity of each CNN model was determined by analyzing the required training time and the Graphics Processing Unit (GPU) memory. In addition, a comparison between the best predictive performance of every CNN was performed. As a result, suggestions could be made regarding the scenarios in which it may be possible to apply specific models, e.g. light-weight CNNs in mobile devices, such as MobileNet [41]. This could lead to a reduction in invasive treatments and the associated economic resources required.

Fifthly, a model-agnostic interpretation tool is used to show the pixels activated during prediction [59]. In this manner, dermatologists can see the probability that a skin lesion is malignant or not, and also which areas are leading the algorithm to give such a conclusion. Nevertheless, bear in mind that this method may not directly generate explanations, and should always be combined with domain-specific knowledge.

Finally, a set of guidelines were given about how to use the studied techniques in order to achieve an accurate CNN. In addition, suggestions were made regarding which areas would be interesting to consider for further study.

The publications associated to this part of the dissertation are:

E. Pérez, O. Reyes, and S. Ventura (2019). ***Diagnóstico Automático de Melanoma: una revisión***. Proceedings of the VII Congreso Científico de Investigadores en Formación de la Universidad de Córdoba (Spain). Available from personal link.

E. Pérez, O. Reyes, and S. Ventura (2021). ***Convolutional neural networks for the automatic diagnosis of melanoma: An extensive experimental study***. Medical Image Analysis, 67. https://doi.org/10.1016/j.media.2020.101858.

The above analysis allows us to develop the following proposals.

4.2. Genetic algorithms to discover novel architectures

In spite of several ensemble-based CNN models proposed in the literature, most of them involve training the models and merging their predictions afterwards. However, CNNs enable more options. For example, each CNN extracts abstract features which are used to obtain a prediction. Such features from all the models could be used to obtain an advanced prediction. In addition, during the training process n sets of weights and predictions are also obtained for each CNN, n being the number of epochs the models are trained. As a consequence, an optimal case could be to obtain those weights, abstract features, and predictions that contribute the most to the ensemble.

With the aim of improving the predictive performance of ensembles, a genetic algorithm for the automatic selection and training process of CNN models is proposed

[J5]. The main aim is to select the CNNs that best contribute to the ensemble, rather than the individual level. In this work a genetic algorithm is used to find the best set of CNNs to build an ensemble-based architecture. This type of approach has previously shown that it can yield a more stable learning, a better convergence [52], as well as a regularizing effect [77]. We extract the abstract features from each CNN, which can represent a different feature space, thus reducing the large number of representations that would be needed to make the model transformation-invariant. Then, an extra prediction block is powered by the above features, which provides an additional prediction. The above predictions were considered to obtain a final one.

On the other hand, a Genetic Algorithm (GA) was designed to learn the optimal sets of CNN, which are used to create ensembles. Each individual represents an ensemble and a full solution of the problem. The population was randomly created, but repeated individuals were not allowed. Each individual was evaluated by training its predictive blocks during 20 epochs and obtaining the average loss value between the chosen CNN models and the extra prediction block; a lower average means a more desirable individual. Regarding evolving the individuals, a custom flat crossover and a mutation operator were used. Finally, the proposal used a generational elitist algorithm, which means that the best individual in the last generation is also the best of all the evolution.

We performed an extensive experimental study, where the number of datasets was increased from 11 to 16. Firstly, we validated a proposed extension of the Chan-Vese segmentation algorithm [15] by using two specialized datasets, which include binary segmentation masks that identify the lesion areas. The proposal is based in specific knowledge from the skin lesion diagnosis task and does not need a previous training step. The segmentation masks indicated that simpler methods could achieve better performance than more complex CNN-based segmentation architectures. Overall, all baseline CNN models achieved the best average performance using segmented images.

Secondly, the advantages of using the extra prediction block were analyzed, which achieved significantly better performance compared to simpler baseline ensembles. In addition, the best ensemble attained in the previous step is used in the next comparisons. Thirdly, the impact of three main hyperparameters was analyzed. This analysis allowed us to shed light on the components that control the GA, as well as to choose the options that could surpass the baseline ensembles. Fourthly, the best architecture was compared to the baseline individual CNN models, which attained the best predictive performance throughout all the datasets. Also, Wilcoxon's test found that the proposed approach achieved significantly better performance than the best baseline ensemble, confirming the benefit and effectiveness of using genetic algorithms. Furthermore, the datasets were classified in dermoscopic and non-dermoscopic. The proposal still attained 13% and 14% better performance than the average comparable architecture considering dermoscopic and non-dermoscopic images, respectively. Finally, a web application[1], where a discovered ensemble-based model is embedded, was developed in order to promote further technological applications in medical diagnosis. This application enables experts to use both main features: lesion segmentation and prediction.

On the other hand, we propose a first approach about a multimodal predictive model. The proposed method combines clinical and imaging data, genetic algorithm, and classical techniques such as data augmentation and transfer learning. In addition, several fusion strategies are analyzed, as well as a search regarding how to combine features and predictions. Experimental results show better performance than an imaging-based CNN model. However, a deep evaluation is encouraged, including a variety of models and more data, whenever possible. On top of that, the proposal is able to distinguish between eight categories, which is a big step towards a more specific diagnosis.

The publications associated to this part of the dissertation are:

E. Pérez, and S. Ventura (2021). ***An ensemble-based convolutional neural network model powered by a genetic algorithm for melanoma diagnosis***. Neural Computing and Applications. https://doi.org/10.1007/s00521-021-06655-7.

E. Pérez, and S. Ventura (2022). ***Multi-view Deep Neural Networks for multiclass skin lesion diagnosis***. Proceedings of IEEE International Conference on Omni-Layer Intelligent Systems. Accepted.

[1] http://skinensemble.com

4.3. Novel imaging methods for skin image analysis

We have demonstrated the suitable performance of ensemble-based models. However, other three methods have been proposed, which are focused on the following promising areas: generation of synthetic skin lesion images, training speed enhancement, and quality of the extracted abstract features.

The first proposal considers the spatial relationships between the extracted features, which could be lost in CNN models [85]. CNNs are approximately invariant to small translations, but they are neither rotation, color nor lighting-invariant [54, 73]. For example, CNN models could consider two images to be similar if they share the same objects, even if the location within the image is relevant. Invariance to a transformation is an important concept in image recognition area, and this means if one takes the input and transforms it, then the representation one gets is the same as the representation of the original. Due to the great variety of possible morphologies of moles, this is an important issue to solve in order to attain a more effective melanoma diagnosis.

Now we aim to design a simpler architecture, which is composed of a single predictive block and uses a recently proposed approach to detect spatial hierarchies between entities within an image [J4]. In addition, a convolution-based computational block is initially used in order to extract abstract features before passing them to the main predictive block. The main block is based on CAPSNET, whose equivariant characteristics make it more attractive since the model is capable of detecting the rotation or proportion change and adapt itself in a way that the objects are internally represented as vectors.

Firstly, the baseline CAPSNET architecture was optimized for the diagnosis of melanoma. In this regard, three optimizers were analyzed, and Stochastic Gradient Descend (SGD) achieved the best average performance. Nevertheless, the architecture was not significantly sensitive to the optimizers. Secondly, we performed a hyper-tuning of the two main components: the dimensions of the primary caps and the class caps. The first ones detect abstract features about position, size, orientation, deformation, among others. The second ones served as the final predictive features, which in this case was 64 features for each class (malignant and benign).

Thirdly, the baseline was boosted by applying data augmentation, increasing the performance between 4% and 78%, and 39% on average.

Fourthly, the full proposal, which includes an advanced convolutional block, was evaluated by following the same experimental settings. The results corroborated that the proposal had achieved significantly better performance than six state-of-the-art CNN models. Regarding transfer learning, the proposed architecture still achieved the best average performance and significantly better performance than four models.

Fifthly, although the results show that our proposed architecture is effective for solving the problematic task of diagnosing melanoma, we encourage the "explanations" of the individual predictions in order to double-check the outputs. As a consequence, Shapley Additive Explanations (SHAP) [59] and Local Interpretable Model-agnostic Explanations (LIME) [69] methods are employed, which are able to highlight the pixels with the highest activation. These methods may not directly generate explanations, but when they are combined with domain knowledge, they can achieve quantitative comparisons and insights in individual predictions. Several lesions were analyzed, which corroborated that the proposal had focused on the main part of the lesions during prediction.

To sum up, the results indicated that the foundation of CAPSNET is only suitable for the diagnosis of melanoma when it is combined with advanced feature extraction techniques in early phases. The above broadens the path to use primary caps in new predictive models, which will be able to identify if some objective element exists and if so what its characteristics. In fact, CAPSNET have been used recently not only in image-based prediction tasks [82], but also in sequence learning [75], which validated such an approach.

The second approach aims at overcoming the limited training data by using a Progressive Growing Generative Adversarial Network (PGGANs) architecture [J6]. Each internal layer of the architecture gets features from at least two previous layers, which is inspired by Dense Convolutional Network. Additional inputs from preceding blocks help to ease the training process. In addition, each architecture is specialized in only one category, i.e. melanoma or nevus. By doing this, the stability of the training process is increased, achieving high-quality 512×512 dermoscopic and non-dermoscopic images, particularly in small datasets. The synthetic skin lesion images could be used for education, which represents an added value. The proposal has been evaluated qualitatively and quantitatively through the use of an extensive experimental study on sixteen dermoscopic and non-dermoscopic skin image datasets (having from 60 to 17393 images), illustrating its effectiveness in the diagnosis of melanoma. Furthermore, the results validate the suitability of the proposal, in which four state-of-the-art data augmentation techniques applied in five convolutional neural network models were significantly outperformed.

The third method proposes a pipeline to train CNN models by analyzing how informative each sample is [J7]. We hypothesize that a better performance in least number of epochs could be achieved if CNN models are able to analyze beforehand from where training is performed. A custom active learning approach guides the training process, where the convolutional architecture is benefited from its uncertainty about individual skin images. Three query strategies and a pool-based scenario are used, selecting which samples go in each mini-batch. As a result, the training process is adjusted after each epoch, achieving a better performance than random selection. The diagnosis is enhanced by using image segmentation, data augmentation and transfer learning. An extensive experimental study was conducted, where five state-of-the-art models were significantly outperformed. The proposal requires 2% of the total training time, and needs 61% less training epochs. Overall, 11% and 20% better predictive performance were achieved in dermoscopic and non-dermoscopic images, respectively.

The publications associated to this part of the dissertation are:

E. Pérez, and S. Ventura (2021). ***Melanoma Recognition by Fusing Convolutional Blocks and Dynamic Routing between Capsules***. Cancers, 13(19). https://doi.org/10.3390/cancers13194974.

E. Pérez, and S. Ventura (2022). ***Progressive growing of Generative Adversarial Networks for improving data augmentation and skin cancer diagnosis.***. Artificial Intelligence In Medicine. Submitted.

E. Pérez, and S. Ventura (2021). ***A framework to build accurate Convolutional Neural Network models for melanoma diagnosis***. Knowledge-Based Systems. Submitted for second revision round.

5

Conclusions and future work

The development of this Ph.D. thesis has several contributions to the diagnosis of melanoma. To begin with, the experimental review allows us to identify as baseline models some of the most used CNN models for the diagnosis of melanoma. Also, the guidelines outlined here serve as a basis for subsequent research. Further contributions of the thesis focus on several aspects of training CNN models. In this matter, Section 5.1 summarizes the conclusions and suggestions that can be drawn from this research. The novel approach taken in this thesis prompts further investigation into how the proposed approaches may be applied to other areas of medical diagnosis, for example in breast and lung cancer. In addition, this Ph.D. thesis could be utilized for boosting the development of cheaper diagnostic tools in order to assist dermatologists in decision making. Moreover, Section 5.2 establishes promising lines of research, which encourage a balance between diagnostic performance and the exploration of low-cost techniques and models.

5.1. Concluding remarks

This Ph.D. thesis has explored the use of several techniques for classifying skin lesion images in the context of early detection. Firstly, we have presented a deep experimental study of CNN models. Secondly, an evolutionary method for creating ensembles of CNN models and a novel architecture which merges convolutional blocks and CAPSNET were presented. The above contributions have been published in three journals. The experimental review is showed in [J3], while the GA to create ensembles is in [J5], and finally the architecture is presented in [J4]. Then, three more papers have been submitted - a multimodal architecture discovered by a genetic algorithm is showed in [66], a Generative Adversarial Network architecture to generate synthetic data is in [J6], and a framework to guide the training process is in [J7].

In addition, we have included two more papers related with this thesis. In [J2], a machine learning methodology for the automatic cancer diagnosis from RNA-Seq data is proposed. Firstly, a ranking of features (isoforms) is determined by using feature weighting algorithms. Secondly, the best subset of features is determined by conducting an effective heuristic search. Thirdly, the confidence of a classifier is assessed by explaining the individual predictions and its global behavior. The results showed that our method achieved better performance than state-of-the-art methods, specifically when identifying skin cutaneous melanoma. Also, the average importance ranking of each gene/isoform is shown, which allows researchers to discover and focus on new interesting behaviors in RNA-Seq data. However, due to the limited amount of public RNA-Seq data regarding skin lesions, we aimed to collect more data in order to gain a broad view of our task.

Specifically, the main contributions that can be highlighted are:

Experimental review. The experimental review on CNN models for the diagnosis of melanoma provides a broad overview of the main techniques and state-of-the-art algorithms. Firstly, CNNs are ranked according to the number of parameters to train, which commonly leads us to identify the possible models to consider in order not to exceed computational resources. Secondly, an analysis of the data is presented, aimed at discovering how complex the current task is and the most demanding datasets. This analysis represents, to the best of our knowledge, the first one to extract individual characteristics from each dataset. Thirdly, the extensive experimental setup has concluded that transfer learning and data augmentation are effective techniques to cope with imbalance and the lack of data. Also, the use of residual connections is recommended, which allows us to build deeper models and mitigates the vanishing-gradient problem. To sum up, overall predictive performance indicated that MobileNet was the best option, closely followed by DenseNet201. In addition, MobileNet was the fastest model, needing only half the time of its nearest competitor - VGG16. However, it is worth mentioning that Xception achieved the top performance in HAM10000, which was the biggest and more recent dataset of skin lesion images at the time.

Genetic algorithm to discover new architectures Evolutionary methods have proven to be effective in the diagnosis of melanoma, as stated in this work. The proposed framework is able to search suitable architectures of ensembles automatically. In addition, the proposal represents a novel approach to finding the abstract features which best contribute within the ensemble and at the same time to training such ensembles of CNNs. The loss-based fitness function merges the performance obtained by the members of the ensemble and the prediction blocks, aimed at achieving a more stable training process. An analysis of the parameters, which guide the GA, also provides useful guidelines to select the final architecture. An experimental study has corroborated that the proposal overcomes six state-of-the-art CNN models in both dermoscopic and non-dermoscopic datasets. Also, the best baseline ensemble is significantly surpassed.

Secondly, a multimodal multiclass neural network architecture, which is capable of overcoming its baseline imaging convolutional model, is discovered by following a simple genetic algorithm. In this manner, we aim at tackling one of the biggest concerns of radiologists - clinical information have a significant impact during diagnosis. Nowadays, most research on skin lesions only use imaging data, mainly because a lack of data. Our first approach feeds from imaging and clinical data, and explores three fusion approaches. Each fusion strategy have its own settings, which is optimized and shared with the others. As a result, suitable parameters can be found with less effort and input data. It is well-known that CNN models require an enormous amount of data in order to achieve an acceptable performance. In this work a total of 72,106 images and meta-data from eight skin lesion categories are considered. Although the method achieved 85%, 94%, and 84% of recall score when diagnosing malignant lesions - melanoma, basal cell carcinoma, and squamous cell carcinoma, respectively, there is a gap of improvement.

Finally, these approaches are not restricted to melanoma diagnosis problems, and could be applied to other real-world medical scenarios, such breast cancer diagnosis.

Analysis of the advantages of merging convolutional blocks and CAPSNET. In this work, a method based on deep feature sharing and enhanced routing is proposed in order to take advantage of both approaches. Firstly, a hyper-tuning of the main parameters is performed, which allows us to adapt the architecture to a bigger input of 299 × 299 × 3 pixels. Bear in mind that CAPSNET was originally trained with 28 × 28 × 1 pixels. Secondly, the capsules were fused with an advanced convolutional block, which is able to extract richer features. A comparison between its use and non-use is also available, showing that overall, the proposal exceeds the baseline method by 23% through using data augmentation techniques. The above suggests that the higher the resolution of the images, the more processing is required. Thirdly, a deep experimental study was carried out, where the proposal significantly outperformed four state-of-the-art models. Fourthly, SHAP and LIME methods were used to highlight the pixels activated during prediction. Combining the above methods and the expertise of dermatologists could lead to finding insights into individual predictions. Furthermore, we have proven the potential of these techniques to detect malignant skin lesions by transferring the convolutional features to the next primary capsules, and finally to the class capsules. Finally, although the architecture was tuned for melanoma diagnosis, it could be applied in other scenarios.

5.2. Future lines of research

The contributions of this Ph.D. thesis have confirmed the usefulness, effectiveness and robustness of different architectures, and optimization techniques in solving the automatic diagnosis of skin lesions. Overall, the proposed approaches are not restricted to the above domain, and can be adapted to other scenarios. In this way, this thesis paves the way for further lines of research that will build upon the main ideas outlined above.

Fusion of medical imaging and electronic heath records

The results achieved in this work showed that several CNN models and techniques are suitable for the diagnosis of melanoma. However, non-imaging data based on electronic heath records enables physicians to interpret imaging outcomes in the right clinical context, leading to a higher diagnostic performance, informative clinical decision making, and improved patient explanations. Several types could be used, such as imaging pixel data, structured laboratory data, unstructured narrative data, audio, among others. In this regard, different data fusion techniques could be applied to combine electronic heath data.

Data fusion refers to the process of joining multiple types of data, aimed at extracting more information in order to achieve a better predictive performance. However, the above technique is rarely used in comparison to single modal. So far, three main fusion strategies can be enumerated when fusing medical imaging and meta data: early fusion, joint fusion (or feature level fusion), and late fusion (or decision-level fusion). The first one is aimed at joining multiple feature vectors into a single vector before feeding into a machine learning model for training. Feature vectors can be extracted from images through using CNNs or handcrafted techniques. On the other hand, electronic heath data could be injected directly as features, or it can first feed feedforward networks, and then output a vector with abstract features. Bear in mind, however, that the vector of abstract features will remain fixed, as well as the models responsible for extracting the vectors, and only the machine learning model, which outputs the predictions, will be optimized. Joint fusion is aimed at merging intermediate features (as is the case for early fusion), but the final loss is propagated back to the feature extraction step, which should improve the whole architecture after every training iteration. Finally, late fusion trains each model independently with a unique type of data, e.g. imaging data. It is noteworthy that the default version of this approach does not use a final predictive block, which is the main difference compared to previous proposals. After that, each model outputs a prediction. Such predictions could be aggregated by using max, mean, and majority voting.

Nowadays, research focuses mostly on early fusion, which is the easiest of the fusion strategies. The idea of training several models simultaneously has achieved suitable performance, as stated in [J5]. Multimodal data should improve the performance and further regulate the training process, which helps to avoid overfitting. A first approach is given in [66], where we propose a simple genetic algorithm in order to find an acceptable imaging-clinical architecture. However, further efforts should aim at adding key data, such as narrative data, in order to be more consistent with the real-life diagnostic environment.

Developing efficient computational methods

Hardware limitations should be also considered when designing the experimental setup, which affects the training time and consequently, the number of possible models to be explored. For example, while Zoph *et al.* [89] was performing architecture searches for NASNet, their workqueue system was revealed, which consisted of 500 × NVidia P100s (12 and 16 GB memory) trained during four days. The previous challenge was addressed in [J5] by focusing the genetic algorithm on the prediction blocks. As a result, the number of parameters to train is kept minimal compared to a full ensemble of CNN models. Nevertheless, we encourage the use of more efficient techniques for a sustainable development of predictive models, such as Active Learning [J1]. For example, these techniques could lead to selecting which instances are harder for the model, and then we would take action in order to mitigate known CNN issues, such as overfitting.

On the other hand, the number of built-in camera devices is increasing sharply, which enables the possibility of accessing an almost unlimited amount of data. In this regard, it is worth noting the size of the proposed models, which should be capable of being handled by such mobile devices. On top of that, nowadays privacy is becoming an important issue, leading us to select models which could even be trained on-device, such as MobileNet. Furthermore, although automatic diagnosis mobile applications exist[12], the adoption of such methods in daily-life scenarios is still limited, as a long-time validation process is required in order to ensure reliable and usable tools. However, most available applications require sending your personal photo to unknown servers, which could represent a clear vulnerability to privacy. Further efforts should be aimed at developing full functional mobile-first models.

[1] https://www.skinvision.com

[2] https://www.firstderm.com/

Design of architectures based on novel methods

Although the baseline CAPSNET is still young, it has been successfully applied in several fields. The proposed architecture in [J4] needs further study in order to take full advantage of the properties extracted from objects in the lesion. This architecture replaces the neuron in a CNN by a group (capsule) of them. Each capsule represents internal properties and is able to learn relationships within objects of the image in order to achieve equivariance. The above solves an issue in fully connected layers, i.e. that they are not able to capture hierarchical structures efficiently to preserve spatial information. Furthermore, it is flexible regarding the design of its blocks. Consequently, custom networks could feasibly be designed, for example by employing another convolutional block with a simpler or more complex internal structure. Moreover, the abstract features extracted from CAPSNET could be used to feed other well-known models, such as Support Vector Machine, which has been to achieve high performance.

Explanation of the outputs

In the last few years, DL models have increased in popularity in the machine learning community, and they are being used for solving a variety of complex problems. Despite its prevalence, the application of DL models is hampered by the fact that black box models obtained with DL are difficult to explain in a comprehensible way. The explanation of the predictions is a challenging and indispensable requirement in order to adopt more seriously automated methods for diagnosing illnesses. This task is crucial for assessing the reliability of the outputs and, as a result, for leading to an effective interaction of the experts with decision support software. For example, biomedicine applications require both an accurate output and a concise explanation [45, 79]. As a result, the experts could verify if the knowledge used by the model matches their own.

In this regard, rule-based models could enhance the utility of DL models by adding interpretability [88], thereby increasing the reliability of the learned knowledge [36]. However, it is noteworthy that extracting rules for DL models cannot be solved in polynomial time (np-hard) [32]. The main challenge could be the development of scalable solutions for complex architectures, since the search space grows exponentially [44]. Furthermore, evolutionary algorithms have proven its effectiveness when extracting knowledge from complex problems [81], and even discovering rules from simple neural networks [43]. The above enables a better exploration of the search space and the generation of accurate rules.

Bibliography

[1] N. R. Abbasi, H. M. Shaw, D. S. Rigel, R. J. Friedman, W. H. McCarthy, I. Osman, A. W. Kopf, and D. Polsky. Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria. *Journal of the American Medical Association*, 292(22):2771–2776, 2004.

[2] W. Abbes and D. Sellami. High-level features for automatic skin lesions neural network based classification. In *Proceedings of the 2nd International Image Processing, Applications and Systems Conference*, Hammamet, Tunisia, 2017.

[3] D. Altamura, S. W. Menzies, G. Argenziano, I. Zalaudek, H. P. Soyer, F. Sera, M. Avramidis, K. DeAmbrosis, M. C. Fargnoli, and K. Peris. Dermatoscopy of basal cell carcinoma: Morphologic variability of global and local features and accuracy of diagnosis. *Journal of the American Academy of Dermatology*, 62(1):67–75, jan 2010.

[4] American Cancer Society. Cancer Facts and Figures, 2020. Consulted on April 14, 2020.

[5] American Cancer Society. Cancer Facts and Figures, 2021. Consulted on June 22, 2021.

[6] American Cancer Society. Cancer Facts and Figures, 2022. Consulted on February 6, 2022.

[7] M. A. Arasi, E.-S. A. El-Dahshan, E.-S. M. El-Horbaty, and A.-B. M. Salem. Malignant melanoma detection based on machine learning techniques: A survey. *Egyptian Computer Science Journal*, 40(03), 2016.

[8] G. Argenziano, S. Puig, I. Zalaudek, F. Sera, R. Corona, M. Alsina, F. Barbato, C. Carrera, G. Ferrara, A. Guilabert, D. Massi, J. A. Moreno-Romero, C. Muñoz-Santos, G. Petrillo, S. Segura, H. P. Soyer, R. Zanchini, and J. Malvehy. Dermoscopy improves accuracy of primary care physicians to triage

lesions suggestive of skin cancer. *Journal of Clinical Oncology*, 24(12):1877–1882, 2006.

[9] U. Asif, M. Bennamoun, and F. A. Sohel. A multi-modal, discriminative and spatially invariant cnn for rgb-d object labeling. *IEEE transactions on pattern analysis and machine intelligence*, 40(9):2051–2065, 2017.

[10] L. Ballerini, R. Fisher, B. Aldridge, and J. Rees. *A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions*, volume 6. 2013.

[11] M. Barzegari, H. Ghaninezhad, P. Mansoori, A. Taheri, Z. S. Naraghi, and M. Asgari. Computer-aided dermoscopy for diagnosis of melanoma. *BMC dermatology*, 5(1):1–4, 2005.

[12] S. Boughorbel, F. Jarray, and M. El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6):e0177678, 2017.

[13] P. Carli et al. Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology. *British Journal of Dermatology*, 148(5):981–984, 2003.

[14] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss. A methodological approach to the classification of dermoscopy images. *Computerized Medical imaging and graphics*, 31(6):362–373, 2007.

[15] T. Chan and L. Vese. An active contour model without edges. In *International Conference on Scale-Space Theories in Computer Vision*, pages 141–151, Corfu, Greece, 1999. Springer.

[16] D. Chicco, N. Tötsch, and G. Jurman. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14(1):13, 2021.

[17] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by ISIC. In *Proceedings of the International Symposium on Biomedical Imaging*, volume 2018-April, pages 168–172, Washington, USA, 2018.

[18] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy. BCN20000: Dermoscopic Lesions in the Wild. 2019.

[19] T. Dietterich. *Ensemble methods in machine learning*, volume 1857 LNCS. 2000.

[20] Ericsson. On the pulse of the networked society. Technical report, 2015.

[21] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[22] F. Félix-Antoine et al. DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13:2171–2175, jul 2012.

[23] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J.-W. W. Coebergh, H. Comber, D. Forman, and F. Bray. Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012. *European journal of cancer*, 49(6):1374–1403, 2013.

[24] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Ann Math Stat*, 11(1):86–92, 1940.

[25] C. Garbe and U. Leiter. Melanoma epidemiology and trends. *Clinics in dermatology*, 27(1):3–9, 2009.

[26] C. Garbe, K. Peris, A. Hauschild, P. Saiag, M. Middleton, A. Spatz, J.-J. Grob, J. Malvehy, J. Newton-Bishop, A. Stratigos, et al. Diagnosis and treatment of melanoma: European consensus-based interdisciplinary guideline. *European journal of cancer*, 46(2):270–283, 2010.

[27] R. Garnavi, M. Aldeen, and J. Bailey. Classification of melanoma lesions using wavelet-based texture analysis. In *2010 International Conference on Digital Image Computing: Techniques and Applications*, pages 75–81. IEEE, 2010.

[28] A. C. Geller, S. M. Swetter, K. Brooks, M.-F. Demierre, and A. L. Yaroch. Screening, early detection, and trends for melanoma: Current status (2000-2006) and future directions. *Journal of the American Academy of Dermatology*, 57(4):555–572, 2007.

[29] S. Gilmore, R. Hofmann-Wellenhof, and H. Soyer. A support vector machine for decision support in melanoma recognition. *Experimental Dermatology*, 19(9):830–835, 2010.

[30] S. Gilmore, R. Hofmann-Wellenhof, and H. P. Soyer. A support vector machine for decision support in melanoma recognition. *Experimental dermatology*, 19(9):830–835, 2010.

[31] I. Giotis, N. Molders, S. Land, M. Biehl, M. Jonkman, and N. Petkov. Med-node: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Systems with Applications*, 42(19):6578–6585, 2015.

[32] M. Golea. On the complexity of rule extraction from neural networks and network querying. In *Rule Extraction From Trained Artificial Neural Networks Workshop, Society For the Study of Artificial Intelligence and Simulation of Behavior Workshop Series (AISB)*, volume 3, 1996.

[33] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[35] D. Gutman, N. C. F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by ISIC. may 2016.

[36] T. Hailesilassie. Rule extraction algorithm for deep neural networks: A review. *arXiv preprint arXiv:1610.05267*, 2016.

[37] B. Harangi et al. Classification of Skin Lesions Using An Ensemble of Deep Neural Networks. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, volume 2018-July, pages 2575–2578, Honolulu, HI, USA, 2018.

[38] K. He et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, NV, USA, 2016.

[39] M. Herland, T. M. Khoshgoftaar, and R. Wald. A review of data mining using big data in health informatics. *Journal of Big data*, 1(1):1–35, 2014.

[40] G. Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386, 1988.

[41] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, page 9, 2017.

[42] L. Jin, S. Gao, Z. Li, and J. Tang. Hand-crafted features or machine learnt features? together they improve RGB-D object recognition. In *Proceedings of the IEEE International Symposium on Multimedia (ISM-2014)*, pages 311–319, Taichung, Taiwan, 2015.

[43] H. Kahramanli and N. Allahverdi. Rule extraction from trained adaptive neural networks using artificial immune systems. *Expert Systems with Applications*, 36(2 PART 1):1513–1522, 2009. cited By 51.

[44] S. Kamada and T. Ichimura. Knowledge extracted from recurrent deep belief network for real time deterministic control. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 825–830. IEEE, 2017.

[45] S. Kamruzzaman, M. Islam, et al. An algorithm to extract rules from artificial neural networks for medical diagnosis problems. *arXiv preprint arXiv:1009.4566*, 2010.

[46] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019.

[47] I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[48] K. Korotkov and R. Garcia. Computerized analysis of pigmented skin lesions: a review. *Artificial intelligence in medicine*, 56(2):69–90, 2012.

[49] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[50] A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 2, pages 1097–1105, Harrahs and Harveys, Lake Tahoe, NV, USA, 2012.

[51] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998.

[52] C.-Y. Lee et al. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570, San Diego, California, USA, 2015.

[53] U. Leiter and C. Garbe. Epidemiology of melanoma and nonmelanoma skin cancerâÄîthe role of sunlight. *Sunlight, vitamin D and skin cancer*, pages 89–103, 2008.

[54] K. Lenc and A. Vedaldi. Understanding Image Representations by Measuring Their Equivariance and Equivalence. *International Journal of Computer Vision*, 127(5):456–476, 2019.

[55] L. Li, Q. Zhang, Y. Ding, H. Jiang, B. H. Thiers, and J. Z. Wang. Automatic diagnosis of melanoma using machine learning methods on a spectroscopic system. *BMC medical imaging*, 14(1):1–12, 2014.

[56] Y. Li and L. Shen. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors (Switzerland)*, 18(2), 2018.

[57] X. Liu, X. Wang, and S. Matwin. Proceedings of the Interpretable Deep Convolutional Neural Networks via Meta-learning. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2018-July, Rio de Janeiro, Brazil, 2018.

[58] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80:14–23, 2015.

[59] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 4765–4774, Long Beach, CA, USA, 2017.

[60] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang. Fusing fine-tuned deep features for skin lesion classification. *Computerized Medical Imaging and Graphics*, 71:19–29, jan 2019.

[61] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, and I. Ellinger. Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. *Computer Methods and Programs in Biomedicine*, 193, 2020.

[62] T. Mendonca, P. Ferreira, J. Marques, A. Marcal, and J. Rozeira. Ph2 - a dermoscopic image database for research and benchmarking. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5437–5440, Osaka, Japan, 2013.

[63] F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig. The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994.

[64] R. Nicolas, A. Fornells, E. Golobardes, G. Corral, S. Puig, and J. Malvehy. Derma: A melanoma diagnosis platform based on collaborative multilabel analog reasoning. *The Scientific World Journal*, 2014, 2014.

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[66] E. Pérez and S. Ventura. Multi-view deep neural networks for multiclass skin lesion diagnosis. 2022.

[67] F. Perez, C. Vasconcelos, S. Avila, and E. Valle. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 303–311. Springer, Granada, Spain, 2018.

[68] O. Reyes, E. Pérez, M. Rodríguez-Hernández, H. Fardoun, and S. Ventura. Jclal: A java framework for active learning. *Journal of Machine Learning Research*, 17:1–5, 2016. cited By 20.

[69] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[70] A. Romero Lopez, X. Giro-I-Nieto, J. Burdick, and O. Marques. Skin lesion classification from dermoscopic images using deep learning techniques. pages 49–54, 2017.

[71] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, nov 1987.

[72] P. Rubegni, M. Burroni, R. Perotti, M. Fimiani, L. Andreassi, G. Cevenini, G. Dell'Eva, and P. Barbini. Digital dermoscopy analysis and artificial neural network for the differentiation of clinically atypical pigmented skin lesions: a retrospective study. *Journal of investigative dermatology*, 119(2):471–474, 2002.

[73] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. *arXiv preprint arXiv:1710.09829*, 2017.

[74] J. P. Shaffer. Modified sequentially rejective multiple test procedures. *J Am Stat Assoc*, 81(395):826–831, 1986.

[75] I. Shahin, N. Hindawi, A. Nassif, A. Alhudhaif, and K. Polat. Novel dual-channel long short-term memory compressed capsule networks for emotion recognition. *Expert Systems with Applications*, 188, 2022.

[76] X. Sun, J. Yang, M. Sun, and K. Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, pages 206–222, Amsterdam, The Netherlands, 2016. Springer.

[77] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[78] C. Szegedy et al. Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June-2015, pages 1–9, Boston, Massachusetts, USA, 2015.

[79] B. J. Taylor and M. A. Darrah. Rule extraction as a formal method for the verification and validation of neural networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 5, pages 2915–2920. IEEE, 2005.

[80] P. Tschandl, C. Rosendahl, and H. Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 2018.

[81] S. Ventura, J. M. Luna, et al. *Pattern mining with evolutionary algorithms.* Springer, 2016.

[82] J. Wang, S. Guo, R. Huang, L. Li, X. Zhang, and L. Jiao. Dual-channel capsule generation adversarial network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022.

[83] J. Wang and L. Perez. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv preprint arXiv:1712.04621*, 2017.

[84] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83, 1945.

[85] C. Xiang, L. Zhang, Y. Tang, W. Zou, and C. Xu. Ms-capsnet: A novel multi-scale capsule network. *IEEE Signal Processing Letters*, 25(12):1850–1854, 2018.

[86] X. Yuan, Z. Yang, G. Zouridakis, and N. Mullani. Svm-based texture classification and application to early melanoma detection. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4775–4778. IEEE, 2006.

[87] X. Zhen et al. Handcrafted vs learned representations for human action recognition. *Image and Vision Computing*, 55:39–41, 2016.

[88] J. R. Zilke, E. Loza Mencía, and F. Janssen. Deepred–rule extraction from deep neural networks. In *International Conference on Discovery Science*, pages 457–473. Springer, 2016.

[89] B. Zoph, V. Vasudevan, J. Shlens, and Q. Le. Learning transferable architectures for scalable image recognition. pages 8697–8710, 2018. cited By 1694.

Scientific publications

[J1] Reyes, O., Pérez, E., del Carmen Rodríguez-Hernández, M., Fardoun, H. M., & Ventura, S. (2016). JCLAL: A Java Framework for Active Learning. *Journal of Machine Learning Research*, 17(95), 15. http://jmlr.org/papers/v17/15-347.html

[J2] Reyes, O., Pérez, E., Luque, R. M., Castaño, J., & Ventura, S. (2020). A supervised machine learning-based methodology for analyzing dysregulation in splicing machinery: An application in cancer diagnosis. *Artificial Intelligence in Medicine*, 108, 101950. https://doi.org/https://doi.org/10.1016/j.artmed.2020.101950.

[J3] Pérez, E., Reyes, O., & Ventura, S. (2021). Convolutional neural networks for the automatic diagnosis of melanoma: An extensive experimental study. *Medical Image Analysis*, 67. https://doi.org/10.1016/j.media.2020.101858.

[J4] Pérez, E., & Ventura, S. (2021). Melanoma Recognition by Fusing Convolutional Blocks and Dynamic Routing between Capsules. *Cancers*, 13(19). https://doi.org/10.3390/cancers13194974.

[J5] Pérez, E., & Ventura, S. (2021). An ensemble-based convolutional neural network model powered by a genetic algorithm for melanoma diagnosis. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-021-06655-7.

[J6] Pérez, E., & Ventura, S. (2021). Revisiting Progressive Growing of Generative Adversarial Networks: a double residual approach for melanoma diagnosis. *Computer Methods and Programs in Biomedicine*, submitted.

[J7] Pérez, E., & Ventura, S. (2021). A framework to build accurate Convolutional Neural Network models for melanoma diagnosis. *Knowledge-Based Systems*, submitted.

Part II

Scientific Publications

6

Compendium of publications

The results of this Ph.D. thesis are supported by several articles: the three journal articles comprising the compendium and two more articles (directly related to the thesis under review); two journal articles as a result of research collaborations (one of them submitted); all of them published (or submitted), to reference journals[1]. Furthermore, two and three conference papers, which have been published or submitted to international and national conferences, respectively.

[1] All journals are ranked at the first quartile according to the Journal Citation Reports.

6.1. Review of Convolutional Neural Network models, experimental study, suggestions and future research lines.

MEDICAL
IMAGE
ANALYSIS
An international journal of computer vision, visualization and image-guided intervention in medicine and biology
MICCAI
ScienceDirect

Title	Convolutional neural networks for the automatic diagnosis of melanoma: an extensive experimental study
Authors	E. Pérez, O. Reyes, S. Ventura
Journal	Medical Image Analysis
Volume	67
Pages	101858
Year	2021
Editorial	Elsevier
DOI	10.1016/j.media.2020.101858

IF (JCR 2020)	8.545
Category	Computer Science - Artificial Intelligence, Interdisciplinary applications, Medicine and Medical Imaging
Positions	13/139 (Q1), 6/112 (Q1), 6/134 (Q1)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Challenge report

Convolutional neural networks for the automatic diagnosis of melanoma: An extensive experimental study

Eduardo Pérez[a,b], Oscar Reyes[b,a], Sebastián Ventura[a,c,b,*]

[a] Knowledge Discovery and Intelligent Systems in Biomedicine Laboratory, Maimónides Biomedical Research Institute of Córdoba, Córdoba, Spain
[b] Department of Computer Science and Numerical Analysis, University of Córdoba, Córdoba, Spain
[c] Department of Information Systems, King Abdulaziz University, Saudi Arabia Kingdom

ARTICLE INFO

Article history:
Received 21 November 2019
Revised 24 August 2020
Accepted 18 September 2020
Available online 13 October 2020

Keywords:
Melanoma diagnosis
Convolutional neural networks
Dermoscopy images
Weigth balancing
Data augmentation
Transfer learning

ABSTRACT

Melanoma is the type of skin cancer with the highest levels of mortality, and it is more dangerous because it can spread to other parts of the body if not caught and treated early. Melanoma diagnosis is a complex task, even for expert dermatologists, mainly due to the great variety of morphologies in moles of patients. Accordingly, the automatic diagnosis of melanoma is a task that poses the challenge of developing efficient computational methods that ease the diagnostic and, therefore, aid dermatologists in decision-making. In this work, an extensive analysis was conducted, aiming at assessing and illustrating the effectiveness of convolutional neural networks in coping with this complex task. To achieve this objective, twelve well-known convolutional network models were evaluated on eleven public image datasets. The experimental study comprised five phases, where first it was analyzed the sensitivity of the models regarding the optimization algorithm used for their training, and then it was analyzed the impact in performance when using different techniques such as cost-sensitive learning, data augmentation and transfer learning. The conducted study confirmed the usefulness, effectiveness and robustness of different convolutional architectures in solving melanoma diagnosis problem. Also, important guidelines to researchers working on this area were provided, easing the selection of both the proper convolutional model and technique according the characteristics of data.

1. Introduction

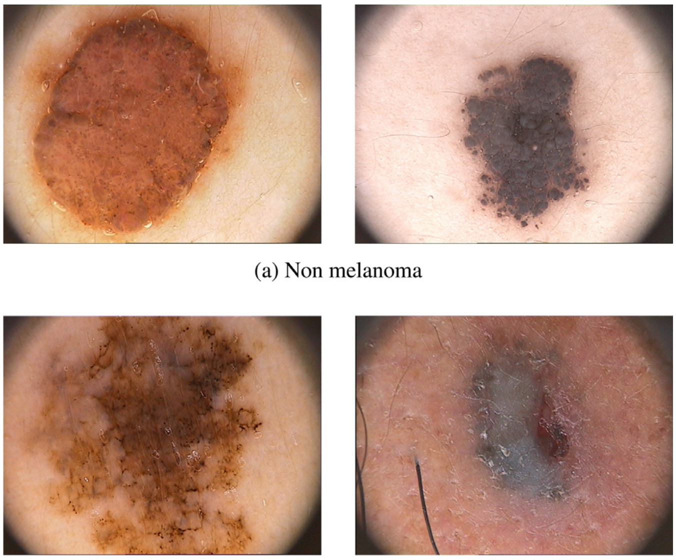
Melanoma is a major public health problem which has the highest levels of mortality among the different types of skin cancer (Siegel et al., 2019). Alarmingly, this illness has an increasing incidence in white people, where just in Europe were estimated 144,200 cases and 20,000 deaths in 2018 (Ferlay et al., 2018), whereas in United States, 100,350 new cases of invasive melanoma and 6850 deaths are expected in 2020 (American Cancer Society, 2020). Consequently, global actions are needed to revitalize efforts for melanoma control and prevention.

To diagnose melanoma, first the dermatologists commonly perform a visual inspection of the skin lesions, and second they conduct a dermoscopic analysis, allowing a better inspection at dermis level (Argenziano et al., 2006; Altamura et al., 2010). However, even after conducting such primary tests, there could still be doubts in the final diagnosis and, finally, a biopsy must be performed. On the other hand, the clinical procedure known as Asymmetry, Border, Color, Diameter, and Evolution (ABCDE) helps in differentiating between normal moles and other types of skin lesions (Nachbar et al., 1994; Abbasi et al., 2004), being possible to determine the stage of melanoma which is directly related to the degree of penetration into the skin. Despite the progress achieved by these clinical tests, the early diagnosis of melanoma remains as a tough task even for expert dermatologists because the complexity, variability and dubiousness of the symptoms (Geller et al., 2007). As a matter of example, Fig. 1 shows that there are a great variety of morphologies in moles of patients having the same biological condition. Accordingly, the melanoma diagnosis problem poses the challenge of developing equally efficient methods that ease the diagnostic and aid dermatologists in decision-making.

Several studies have shown that the early diagnosis of melanoma can be benefited from computational methods (Lee et al., 2018), and recent works demonstrated that such techniques may even overcome the diagnosis made by expert dermatologists (Haenssle et al., 2018). Over the years, several machine learning techniques have been applied for the automatic diagnosis of melanoma from dermoscopic images, e.g. methods based on k-nearest neighbors approach (Li et al., 2014), support vector machines (Gilmore et al., 2010) and random forest (Rastgoo et al.,

[*] Corresponding author.
E-mail addresses: eduardo.perez@imibic.org (E. Pérez), ogreyes@uco.es (O. Reyes), sventura@uco.es (S. Ventura).

https://doi.org/10.1016/j.media.2020.101858
1361-8415/© 2020 Elsevier B.V. All rights reserved.



(a) Non melanoma

(b) Melanoma

Fig. 1. Morphological differences in images of moles belonging to patients which have the same biological condition; images taken from PH2 dataset (Mendonca et al., 2013).

2016). These machine learning algorithms require the previous extraction of handcrafted features, thus incorporating a priori an important knowledge in the analysis, and not requiring the availability of very large datasets for constructing a proper predictive model. However, despite the advantages of these techniques, it should be noted that the quality of the extracted descriptors heavily relies on the level of dermatologists' expertise, and finding informative and discriminative sets of high-level features to build an accurate model remains as a complex and costly task that is usually problem dependent (Jin et al., 2015). In addition, an explicit definition of an intuitive and interpretable high-level feature is sometime impossible to attain because features are derived from pixel intensity space and, therefore, they are not invariant regarding the differences in the input images (Liu et al., 2018). As a consequence of the above, in the last years there has been an increasing attention in developing computational techniques which can automatically extract and learn high-level features, thus providing a higher robustness to the inter- and intra-class variability present in melanoma images (Abbes and Sellami, 2017; Esteva et al., 2017).

In this regard, deep learning models (Reyes and Ventura, 2019b), specifically Convolutional Neural Networks (CNNs), are widely being used for melanoma diagnosis from dermoscopic images (Hu et al., 2018); e.g. an increasing number of contestants in the well-known annual challenge developed by the *International Skin Imaging Collaboration* project (ISIC) are adopting deep learning-based techniques. This type of learning method has the capacity of automatically learning high-level features from raw images (Esteva et al., 2017), allowing the extraction of hierarchies of features by applying convolutional operators that progressively learn more abstract features and, finally, enabling the learning of data-driven features for specific tasks (Zhen et al., 2016). Consequently, CNNs has shown to be more effective in diagnosing melanoma, easing the development of novel applications in a shorter time; e.g Nasr-Esfahani et al. (2016) showed that CNN models can overcome handcrafted features-based methods, and recently Brinker et al. (2019) demonstrated that CNN models can reach prediction levels on par with 145 dermatologists.

Despite the proved effectiveness of CNNs, they are still limited in diagnosing melanoma mainly because need large amount of data to build accurate models (Menegola et al., 2017). In this regard, it is worth noting that most of the existing melanoma datasets only encompass a few hundred of images; although larger datasets have recently appeared (Esteva et al., 2017; Tschandl et al., 2018). Furthermore, in the last years, explaining predictions has become particularly relevant, being crucial for building trust in experts and, therefore, to achieve an effective interaction with machine learning systems (Reyes et al., 2019). CNNs, however, are black-box models and, therefore, there is a lack of explanation regarding how these models reach the final predictions, and this feature could imply both practical and ethical issues in biomedicine (Guidotti et al., 2018). On the other hand, the existing works applying CNNs are commonly limited to the application of a single model that is evaluated over a specific dataset, thus not deeply investigating how to overcome the challenges related to melanoma diagnosis; e.g. the sizeable class imbalance ratio that commonly exists in melanoma datasets (Li and Shen, 2018; Haenssle et al., 2018). Consequently, to date, the understanding of the effectiveness of CNNs for diagnosing melanoma is quite limited, hampering the selection of the most suitable architecture according to the characteristics of data.

As consequence of the above, this work aimed to perform an experimental review focusing on the application of CNNs for diagnosing melanoma. To achieve this objective, an extensive experimentation was conducted, where twelve well-known CNN models were assessed on eleven datasets; ten datasets of dermoscopic images taken from ISIC repository, and one dataset comprising non-dermoscopic images shot with common digital cameras. In the experimental study, first it was analyzed the effectiveness of three well-known optimization algorithms in training the CNNs. Second, two well-known techniques commonly used to deal with imbalanced datasets were evaluated, namely weight balancing (He and Garcia, 2008) (it modifies the cost function by including the weights of each training sample and class group) and data augmentation (Perez and Wang, 2017) (it expands the dataset by applying several random transformations on the original images). Finally, the effectiveness of transfer learning techniques (Weiss et al., 2016) in melanoma diagnosis was also assessed; in this case, the weights learned from the popular ImageNet dataset (Krizhevsky et al., 2012) were transferred to the problem studied. As a result of this work, important insights for researchers working on this area of study are given, thus providing useful guidelines that ease the selection of the proper CNN architecture according the characteristics of data, and also showing the effectiveness of several techniques that can significantly improve the predictive performance in diagnosing melanoma.

The rest of this work is arranged as follows: Section 2 briefly presents several CNN architectures, mainly focusing on those ones that have previously been used in melanoma diagnosis; Section 3 discusses some relevant advanced techniques that may be used for a better training of CNN models and, therefore, to improve their accuracy. Section 4 describes the most popular image datasets for melanoma diagnosis available on the internet, where several characteristics that demonstrate the complexity of this classification problem were analyzed; the experimental design and settings, as well as an analysis of the results and some suggestions are provided in Section 5; finally, Section 6 presents some concluding remarks.

2. Convolutional models and their applications in melanoma diagnosis

In simple words, CNNs are a type of feedforward neural network that use a mathematical operation called convolution, allow-

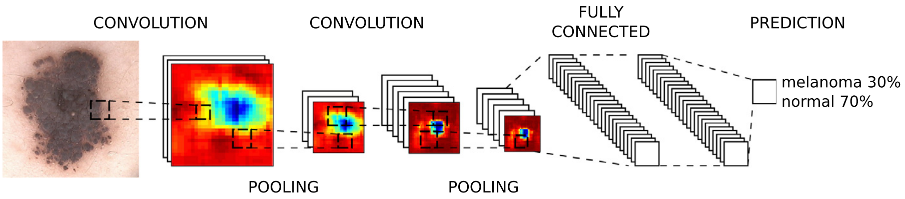


Fig. 2. Melanoma diagnosis with a simple CNN model.

ing to automatically extract features from images, and these features are passed to successive layers that in turn learn more abstract features from previous ones until a final output is yielded, so simulating some of the actions produced in the human visual cortex (LeCun et al., 1998). Consequently, CNNs can build more complex concepts from simpler ones, e.g. they can learn to detect a human face based on simpler concepts as a nose and mouth, which in turn are learned from much simpler ones such as corners and contours (Goodfellow et al., 2016). As a matter of example, Fig. 2 illustrates a simple CNN model that learns to predict whether or not a patient has melanoma by means of continuously mapping the extracted features into more abstract features spaces.

Several CNN models have been proposed since LeCun et al. (1998) presented LeNet in 1998, but their popularity truly increased when AlexNet (Krizhevsky et al., 2012) won the well-known *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) in 2012, reducing the top-5 labels error rate from 26.1% to 15.3%. AlexNet had a very similar architecture as LeNet but it encompassed a larger number of stacked convolutional layers and filters per layer. Later, the ILSVRC-2013 winner was another CNN model named ZFNet (Zeiler and Fergus, 2014), which is mostly an improvement of AlexNet that tweaked its hyper-parameters for reducing the bottlenecks while almost maintaining the same structure. In view of such an effectiveness of the CNNs in solving complex tasks, in the last years new CNN architectures have continuously been appearing, as it can be seen in Fig. 3. It should be noted that most of these CNN architectures follow a modular design, where a building block (which is commonly composed by several stacked convolutions layers) is repeated one after another several times along the network allowing the continuous extraction of more abstract features, and pooling layers are also used between these computational blocks to reduce the feature space, the number of parameters to learn, and also to control overfitting (Goodfellow et al., 2016).

As a consequence of the above, not surprisingly CNNs are actively being used nowadays for melanoma diagnosis (Hosseinzadeh Kassani and Hosseinzadeh Kassani, 2019), bringing new possibilities to reach a higher effectiveness in solving this problem. Evidence of this can be seen in the two last editions of ISIC (ISIC-2018 and ISIC-2019), where the majority of contestant submitted CNN-based solutions. Next, the principal baseline CNN architectures that have been applied (or could be applied) in melanoma diagnosis are summarized, mainly focusing in those ones included and compared in the experimental study carried out in this work.

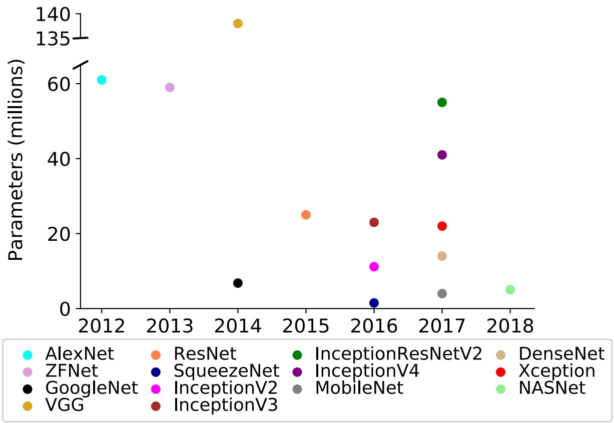


Fig. 3. Main CNN architectures proposed since AlexNet appeared in 2012. Y-axis represents the number of trainable parameters (in millions) of each CNN model.

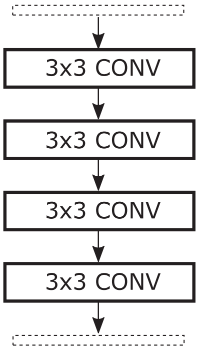


Fig. 4. Building block used in VGG architecture.

VGG architecture

VGG architecture (Simonyan and Zisserman, 2014) demonstrated that deeper CNN models can achieve a better effectiveness, although this type of model requires that a much larger number of parameters must be learned (see Fig. 3). VGG is mainly composed by blocks as the one shown in Fig. 4, where each computational block has four stacked convolutional layers. Several version of VGG can be encounter in the literature, e.g. VGG16 and VGG19, mainly differing in the number of blocks used along the network and filter size applied in each convolutional layer composing a block.

Regarding the application of VGG in melanoma diagnosis, Nasr-Esfahani et al. (2016) proposed a reduced version of VGG by only using two convolutional layers, and obtained results that overcame the predictive performance of several classic state-of-the-art methods. Also, Jaworek-Korjakowska et al. (2019) proposed a computer-vision tool that employs VGG19 for predicting the degree of penetration of skin lesions into one of three possible classes: < 0.75 mm, 0.76-1.5 mm, and > 1.5 mm.

Inception architecture

Inception architecture appeared for the first time at ILSVRC-2014, where GoogleNet (a.k.a. InceptionV1) (Szegedy et al., 2015) won this competition. This architecture uses a block (dubbed as inception module) as the one shown in Fig. 5, where several independent "chains" of convolutional layers with small filters are simultaneously applied over the same input allowing the extraction of more information, and finally the resulting volumes of these chains are concatenated to form a unique representation, which

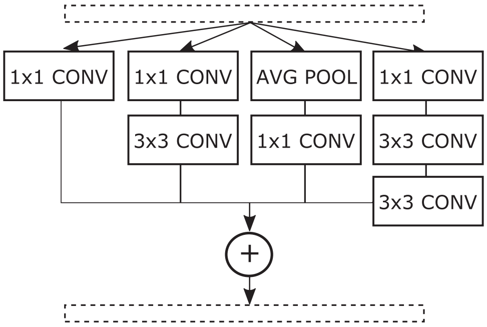


Fig. 5. Building block used in InceptionV3 architecture.

is subsequently passed as input to another inception block. Inception module commonly uses at the beginning of each chain a pointwise convolution (filters of size 1 × 1) to reduce the computational complexity and the number of parameters to learn; this is the main reason that GoogleNet drastically reduced the number of parameters from 60 millions (AlexNet) to 4 millions (see Fig. 3). Later Inception architecture have been redefined in (Szegedy et al., 2016) and (Szegedy et al., 2017), where InceptionV2, InceptionV3 and InceptionV4 were presented. On the one hand, InceptionV2 and InceptionV3 mainly improved the original inception module by means of applying batch normalization technique (Ioffe and Szegedy, 2015), factorized convolutions, and more sophisticated auxiliary classifiers, demonstrating that aggressive dimension reductions can result in networks with relatively low computational cost while maintaining high quality. On the other hand, InceptionV4 followed a more uniform architecture that uses more blocks than InceptionV3, and the authors also proposed a method that makes the model more scalable with memory optimization to backpropagation. Finally, it should be noted that Inception architecture has demonstrated to be useful in contexts where large-scale datasets need to be processed in a reasonable time, as well as when computational resources are limited, such as in big-data problems and mobile applications (Gogul and Kumar, 2017; Chin et al., 2017; Lin et al., 2018).

Regarding the application of Inception in melanoma diagnosis, Esteva et al. (2017) performed a study where a InceptionV3 was used and compared with 21 dermatologists, demonstrating that the model was capable of classifying skin cancer with a level of competence comparable to experts. In this same regard, Haenssle et al. (2018) recently conducted a study using InceptionV4 and it was compared with 58 dermatologists, and they observed that the CNN network obtained better results in melanoma diagnosis than the most of the consulted experts.

ResNet architecture

ResNet architecture was proposed by He et al. (2016), and this model won the ILSVRC-2015 competition. This CNN architecture emerged while studying whether indeed the learning of neural networks is always better by adding more layers. This is because when the number of layers increases, not only the notorious problem of vanishing/exploding gradients appears (this problem, however, can be mitigated by using a normalized initialization), but also accuracy can get saturated and then degrade rapidly (the deep model leads to higher training errors), indicating that not all systems are similarly easy to optimize by just adding more layers.

To solve the aforementioned problem, ResNet introduced the so-called *residual connections* (see Fig. 6a) that directly transfer the input of the block to the output. Residual connections allow that deeper blocks will have the chance to work with non-preprocessed information, and therefore, the added layers can also be constructed as identity mappings, and finally, the model can produce a training error no greater than its shallower counter-part. On the other hand, it should be noted that ResNet is a suitable framework which has commonly been combined with other CNN architectures, like Inception. In this sense, Szegedy et al. (2017) proposed InceptionResNet that exploits the best of both architectures (see Fig. 6b); the authors presented two different versions (InceptionResNetV1 and InceptionResNetV2) that mainly differ in the number of filters, where InceptionResNetV1 roughly has the computational cost of InceptionV3, and InceptionResNetV2 matches the raw cost of InceptionV4.

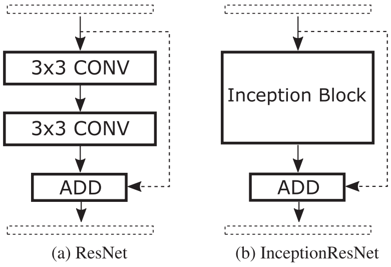


Fig. 6. Building blocks used in ResNet and InceptionResNet.

Regarding the application of ResNet in melanoma diagnosis, Li and Shen (2018) improved the original block used by ResNet by means of incorporating more internal convolutional layers and an extra residual connection, and the proposed model was assessed in ISIC-2017 challenge, obtaining results comparable to the winner of the contest. Han et al. (2018) compared the performance of ResNet152 (it uses 152 convolutional layers) with 16 dermatologists, confirming that the CNN model classified 12 skin tumor types with a prediction level comparable to the consulted experts, whereas the model also showed a superior performance in diagnosing squamous cell carcinoma and melanocytic nevus. Hagerty et al. (2019) recently combined the representations learned by ResNet50 and several clinical features (age, gender, lesion location, etc) to attain better prediction levels in melanoma diagnosis. As for InceptionResNet, despite the effectiveness this combination has shown, it is noteworthy that none or very few studies have considered this type of hybrid model in melanoma diagnosis.

DenseNet architecture

Dense Convolutional Network (DenseNet) (Huang et al., 2017) extends ResNet (see Fig. 7), where each building block receives additional inputs from all preceding blocks and its feature-maps are also passed to all subsequent blocks. Unlike ResNet that performs an element wise addition of the representations, DenseNet concatenates the inputs and outputs by channels. The main module presented in Fig. 7 is then repeated multiple times along the network, and they are connected by transition components composed by convolutional and pooling layers that rescale the outputs to the required size. DenseNet has several advantages, including a mitigation of the vanishing-gradient problem, propagation and reuse

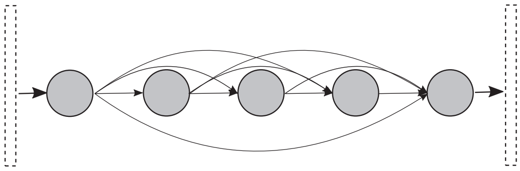


Fig. 7. Building block used in DenseNet.

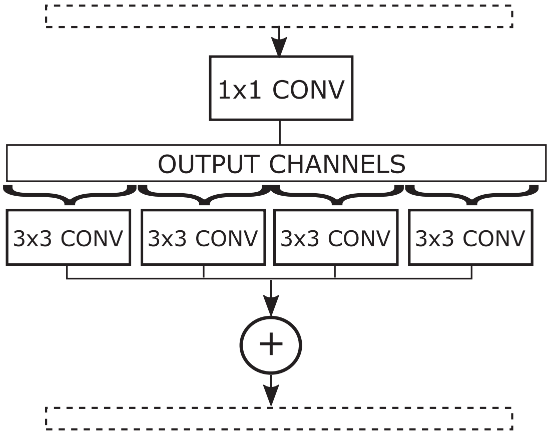


Fig. 8. Building block used in Xception.

of features, and a reduction of the number of parameters with respect to ResNet.

DenseNet has also been applied in melanoma diagnosis, for example Zeng and Zheng (2018) proposed a model named *multi-scale fully convolutional DenseNet* which can perform semantic segmentation of a dermoscopic image with arbitrary size. They proved the model in the dataset released in ISIC-2017 challenge, and obtained promising results.

Xception architecture

Extreme inception (Xception) (Chollet, 2017) is inspired in the popular Inception architecture, but it uses depthwise separable convolution layers, reducing the computational cost. Commonly, this type of layer first performs a depthwise convolution (i.e. various filters performing over each channel of the input independently), and then a pointwise convolution is conducted to get an output with the desired depth. However, Xception proposed to invert the operations, as can be seen in Fig. 8, hypothesizing that the mapping of cross-channels correlations and spatial correlations in the feature maps can be entirely decoupled. Chollet (2017) observed that Xception had a similar predictive performance to InceptionV3 on ImageNet dataset, although the results showed the former can overcame the later in other datasets.

Few studies have applied this type of model to melanoma diagnosis problem. To the best of our knowledge, the first was Zhao et al. (2019) that proposed a Xception model as a risk degree classifier, allowing to identify patients who are at risk and alert them to go to hospital for further examination. The authors compared the model with 20 professional dermatologists, concluding the proposed classifier outperformed consulted experts.

MobileNet architecture

MobileNet (Howard et al., 2017) was designed to be used in mobiles and applications of embedded vision. This architecture is mainly characterized by an extensive use of depthwise separable convolutions, thus significantly reducing the model size and complexity for training. MobileNet also introduces two hyperparameters that balance the latency and accuracy of the model, namely *width* and *resolution* multipliers; *width multiplier* controls the input width of a layer, whereas *resolution multiplier* controls the input image resolution of the network.

As for diagnosing melanoma, Sahu et al. (2018) built an effective hand-held assistant based on smart-phone and Raspberry-Pi that can classify with a good accuracy level skin lesion images. To improve the accuracy of classification, the authors proposed a hybrid approach that used a pre-trained MobileNet combined with domain-specific features that dermatologists commonly use for inspection purpose, such as the texture of the lesion boundary, the symmetry of the mole, and the boundary characteristics of the region of interest.

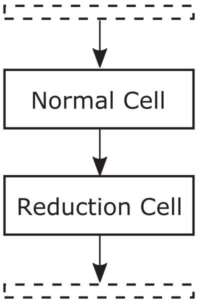


Fig. 9. Building block used in NASNet.

NASNet architecture

Neural Architecture Search (NASNet) (Zoph et al., 2018) is a modern CNN architecture proposed by GoogleBrain that searches an effective building block on a small dataset using a reinforcement learning search method, and transfers the block to a larger dataset, allowing to achieve state-of-the-art results with smaller model size and lower complexity; NASNet applies the block found in the small dataset to the larger dataset by stacking together more copies of this. Instead of designing the building block using hand-crafted decisions, as the aforementioned CNN architectures do, NASNet tries to find cells that can form a block with the best performance, and the internal structure of these cells are determined through a Recurrent Neural Network. Specifically, two types of cells are constructed, namely Normal and Reduction cells (see Fig. 9), and then the building block is repeated n times; this last parameter is automatically determined. Normal cells are comprised by convolutional layers that return a feature map of the same dimension as the input, whereas Reduction cells encompass convolutional layers returning a feature map reduced by a factor of two.

Despite the proved effectiveness of NASNet in solving complex problems (Kim et al., 2019), to the best of our knowledge, this CNN architecture has not previously been applied to melanoma diagnosis.

3. Techniques to improve convolutional neural networks for melanoma diagnosis

In the last few years, there has been an increasing tendency not only in developing and using modern CNN architectures to solve real-world complex problems, but also in applying advanced techniques for achieving a better training of these models, such as transfer learning techniques (Khan et al., 2019), data augmentation methods (Perez et al., 2018), and the development of ensembles of CNNs (Matsunaga et al., 2017) and multi-task models (Twinanda et al., 2017).

Transfer learning is a technique widely used in image classification tasks (Schwarz et al., 2015; Sa et al., 2016; Shin et al., 2016), where CNNs commonly requires a large collection of data in order to build an accurate model. As its name suggests, this type of method tries to transfer and reuse a knowledge that was extracted from a source task, where a lot of data is commonly available, on a target task. Several authors have shown the usefulness of transfer learning in melanoma diagnosis. For example, Khan et al. (2019) employed transfer learning with ResNet50 and ResNet101, and extracted the best learned features that were subsequently used to build a Support Vector Machine (SVM) model. Thao and Quang (2017) used a VGG16 model pre-trained on the well-known ImageNet dataset to perform two different tasks, namely lesion segmentation and classification of skin tumors from dermoscopic images, and they obtained promising results. Huang et al. (2019) also applied a pre-trained VGG16 for skin lesion segmentation, obtaining results that were competitive against several segmentation algorithms. Esteva et al. (2017) employed a pre-trained InceptionV3 and they obtained a performance on par the tested experts, proving the capability of CNNs to help dermatologists. Finally, Yoon et al. (2019) focused on training generalizable data-driven models for skin lesion applications using seven domains and seven datasets, showing that it is better to use data from very similar domains in order to increase the chances of finding generalizable features.

Handling imbalance datasets is another problem that needs to be addressed when training machine learning models, and deep learning ones are not exception. In the melanoma diagnosis task, it is very common to have an uneven distribution of samples per class, where the number of normal samples is several times higher than the number of tumor ones. **Weight balancing, data augmentation** (Goodfellow et al., 2016) and recently **Generative Adversarial Networks** (GANs) (Goodfellow et al., 2014) are three useful techniques that can cope with this issue. The former is a cost-sensitive learning method that balances data by altering the weight that the samples carries when computing the loss, e.g. melanoma images could have a weight (importance) more times higher than normal images while training a deep model (Ling and Sheng, 2010). Data augmentation, on the other hand, tries to augment a dataset by applying random transformations on the original images; also, it is noteworthy data augmentation can be applied to improve the generalization capacity of deep models and as regularization method to prevent overfitting (Ciresan et al., 2010). For example, Perez et al. (2018) demonstrated the positive impact of doing several transformations over the original images for a better melanoma diagnosis, such as color, geometric and elastic transformations, and random erasing. Also, Esteva et al. (2017) augmented a dataset 720 times by means of rotating, cropping, and flipping the original images, achieving better prediction levels. Finally, GANs can augment a dataset by training simultaneously two models: a generative model that captures the data distribution, and a discriminative model that predicts whether a sample is a fake or not. For example, Baur et al. (2018) applied GANs to generate realistically looking high resolution images of skin lesions. The authors proposed a new GAN model based on two classical GANs architectures, namely Deep Convolutional GANs (Radford et al., 2015) and Laplacian Pyramid of Adversarial Networks (Denton et al., 2015), and they showed that this type of method are able to mimic the data distribution with diverse and realistic samples, even when the training dataset is very small.

Ensemble learning is another paradigm that has shown to be very effective for improving the effectiveness of the CNN models in melanoma diagnosis; ensemble methods try to combine the outputs of several weaker classifiers in order to obtain superior results (Dietterich, 2000). For example, Harangi et al. (2018) proposed an ensemble composed by AlexNet, VGG and InceptionV1, and they evaluated their proposal on the IEEE International Symposium on Biomedical Imaging (ISBI) 2017 challenge. Mahbod et al. (2019) performed skin lesion classification combining several CNNs of different architectures, and the sets of partial predictions were aggregated to obtain the final ones. Liberman et al. (2018) also proposed an ensemble composed by three classifiers, a pre-trained VGG16 model and two SVMs that used as input the features extracted by the former.

Finally, multi-task learning (MTL) (Caruana, 1997) is an approach to inductive transfer that improves generalization by learning tasks in parallel while using a shared representation; what is learned for each task can help in solving the other tasks (Reyes and Ventura, 2019a). As for using MTL paradigm in melanoma diagnosis, Kawahara et al. (2019) proposed a multi-task deep CNN that was not only trained with melanoma images, but also with patient meta-data. A pre-trained InceptionV3 was used as base model and the training images were also real-time augmented using flips, rotations, zooms, and shifts. The model was trained using five loss functions, where each loss considered different combinations of the input modalities.

Despite the advantages that some techniques offer for a better training of CNN models, such as GANs, ensemble and multi-task learning methods, it should be stressed that important limitations arise when they are applied. For example, GAN-based models require a complex training process where a balance between the generative and discriminative models needs to be found. Therefore, GANs frequently fail to converge (Berthelot et al., 2017), where gradients can result in random oscillations. Also, setting up a GAN-based model involves a very costly tuning process, which can quite limit its correct application in some real-world problems. As for ensemble models, the training of this type of model commonly requires the assessment of a high number of possibilities that hamper the process, such as the number and type of CNN models to combine, and the way to combine the learned representations and partial predictions yielded by each member of an ensemble. Finally, regarding multi-task learning, the main limitation when applying this paradigm is the lack of public datasets that contain heterogeneous information for each sample (e.g. dermoscopic images with their associated metadata and clinic data), thus limiting quite a lot the development of accurate multi-task models for melanoma diagnosing.

Considering the above, and that the main objective of this work is to assess the suitability of the CNN models in melanoma diagnosis, the experimental study conducted only covers the analysis of the CNN architectures described in Section 2 (as baseline models). How much to improve the predictive performance of the baseline models by means of applying some classical methods was also studied, such as weight balancing, data augmentation and transfer learning.

4. Image datasets for melanoma diagnosis

Due to the incidence of melanoma is continuing to increase worldwide, in the last few years several private and public datasets have been published, thus allowing a better study of this illness and, therefore, the design of better approaches for its automatic diagnosis. The most popular private data collections of dermoscopic images are the *Interactive Atlas of Dermoscopy* (Argenziano et al., 2004), *Dermofit Image Library* (Ballerini et al., 2013), and the dataset presented by Esteva et al. (2017) which conducted a comparison with 21 dermatologists using 129,450 clinical images; to the best of our knowledge this last dataset is the largest one reported in the literature.

Regarding public datasets for studying melanoma, the biggest collection of datasets can be found in ISIC repository[1] which comprises images labeled by expert dermatologists. In this repository we can found the HAM10000, MSK and UDA datasets which appeared in (Tschandl et al., 2018), (Codella et al., 2018), and (Gutman et al., 2016), respectively. Furthermore, this repository provides the different datasets presented in the annual ISIC challenges (ISIC-2016, ISIC-2017, etc), which are also commonly used as benchmarks by the researchers. Mendonca et al. (2013), on the other hand, created the PH2 dataset[2] that comprises 200 high-

[1] https://isic-archive.com

[2] https://www.fc.up.pt/addi/ph2%20database.html

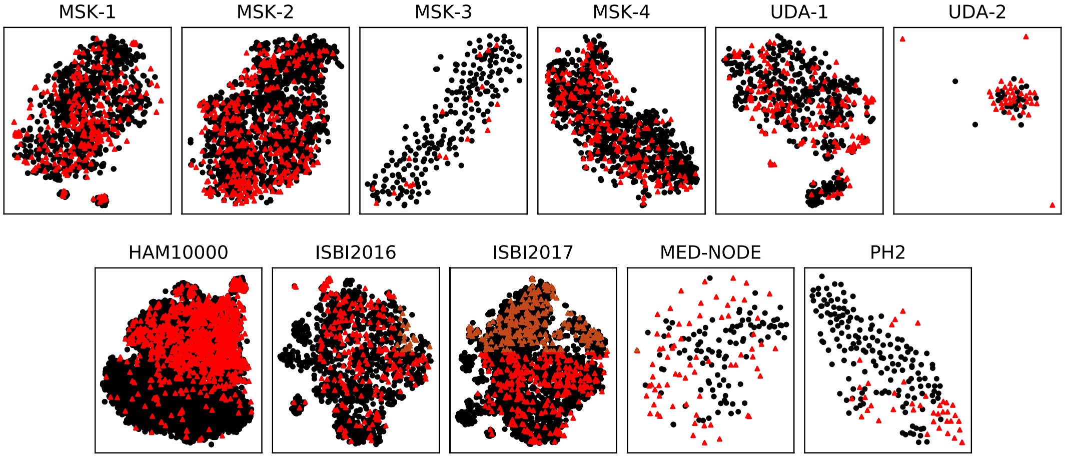


Fig. 10. Visualization of the selected datasets by using t-SNE. Black and red dots represent normal and melanoma samples, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Summary of the benchmark datasets. Img.: total number of images; ImbR.: imbalance ratio between the benign and malignant classes; IntraC.: average distance between samples of the same class; InterC.: average distance between samples of different classes; DistR.: ratio between IntraC. and InterC.; Silhouette: silhouette score.

Dataset	Source	# Img.	ImbR	IntraC	InterC	DistR	Silhouette
HAM10000	(Tschandl et al., 2018)	7818	6.024	8705	9770	0.891	0.213
ISBI2016	(Gutman et al., 2016)	1,273	4.092	10,553	10,992	0.960	0.101
ISBI2017	(Codella et al., 2018)	2745	4.259	9280	9674	0.959	0.089
MED-NODE	(Giotis et al., 2015)	170	1.429	9029	9513	0.949	0.068
MSK-1	(Codella et al., 2018)	1088	2.615	11,753	14,068	0.835	0.173
MSK-2	(Codella et al., 2018)	1522	3.299	9288	9418	0.986	0.062
MSK-3	(Codella et al., 2018)	225	10.842	8075	8074	1.000	0.112
MSK-4	(Codella et al., 2018)	943	3.366	6930	7162	0.968	0.065
PH2	(Mendonca et al., 2013)	200	4.000	12,688	14,928	0.850	0.210
UDA-1	(Gutman et al., 2016)	557	2.503	11,730	12,243	0.958	0.083
UDA-2	(Gutman et al., 2016)	60	1.609	11,297	11,601	0.974	0.020
Average		-	4.003	9939	10,677	0.939	0.109

quality dermoscopic images. Finally, Giotis et al. (2015) presented the MED-NODE dataset[3] that collects 170 non-dermoscopic images shot with common digital cameras; it is noteworthy that this source of data is very important nowadays due to the use of technological devices (e.g. smartphones and tablets) is constantly growing.

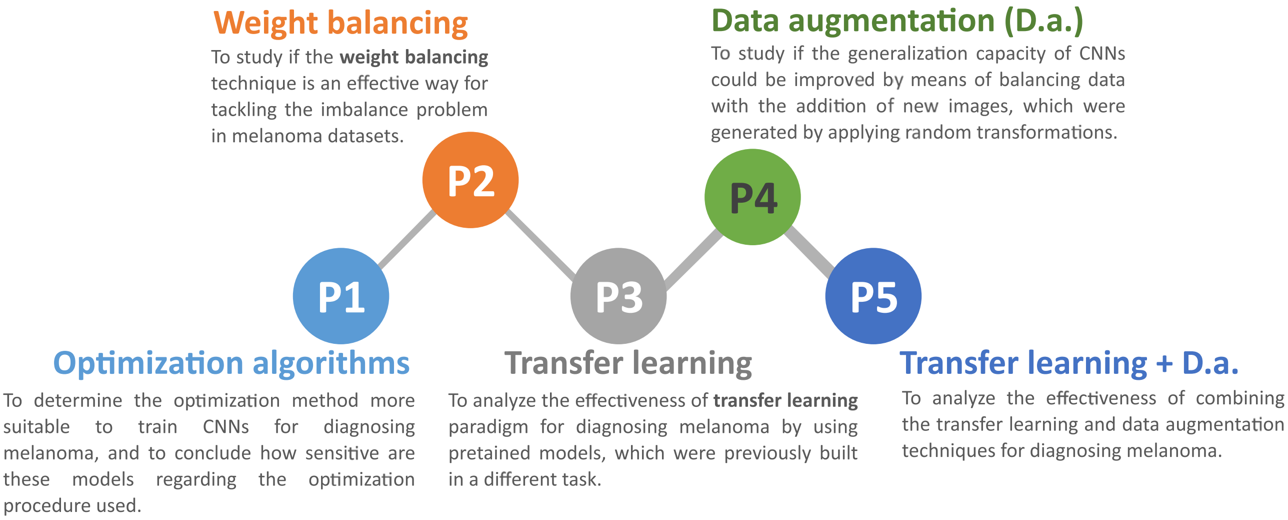
In this work, eleven public datasets were studied. Nine datasets from ISIC archive, the PH2 and MED-NODE datasets were considered in the study, finally processing a total of 16,601 images; each RGB image i had a resolution (h, w, c), being h=224, w=224, and c=3 the height, width, and number of channels of i, respectively. All the datasets encompass binary classification problems, where each image was labeled as benign or malignant sample. Table 1 shows a summary of the characteristics of the studied datasets, where it can be observed that they vary in sizes and complexity; the metrics show the complexity of the underlying classification problem of each dataset. It can be observed that some datasets present a moderate imbalance ratio, indicating that the number of benign samples is several orders of magnitude higher than the number of malignant samples. Also, the existing intra-class and inter-class distances in each dataset were computed using the Euclidean function distance, where each image i was represented as a vector $\mathbf{X}_i$ of length l=$h \times w \times c$. The results showed high values in both intra-class and inter-class distances, indicating that the average distances between images belonging to different classes, as well as between images belonging to the same class, are quite large. Also, the ratio between the intra-class and inter-class distances showed that both distances are similar, and this value commonly indicates a high degree of overlapping between the different class spaces. Finally, the silhouette score (Rousseeuw, 1987) was also measured, which represents how similar an image is to its own cluster compared to other clusters. The low values of Silhouette score indicated that the images were not well matched to their own cluster, and even samples belonging to different clusters are close in the feature space.

Finally, all the selected datasets were visualized by using the well-known *t-distributed Stochastic Neighbor Embedding* (t-SNE) (Laurens van der Maaten and Hinton, 2008), which is a well-suited method for visualizing high-dimensional data; the vectors of RGB values representing each image were projected into a two-dimensional space. Fig. 10 shows that there was a high overlapping level between images belonging to different classes in the projected space (excepting the case of PH2 dataset where a more clear separation was observed), thus indicating the high variability presents in melanoma images and, therefore, confirming the high complexity of the melanoma diagnosis problem.

5. Analysis of the effectiveness of convolutional models for melanoma diagnosis

This section summarizes the extensive experimental study conducted, aiming to analyze the effectiveness of CNN models for melanoma diagnosis. First, the experimental protocol and settings used throughout the study are described. Then, the experimen-

[3] http://www.cs.rug.nl/imaging/databases/melanoma_naevi/



Weight balancing
To study if the **weight balancing** technique is an effective way for tackling the imbalance problem in melanoma datasets.

Data augmentation (D.a.)
To study if the generalization capacity of CNNs could be improved by means of balancing data with the addition of new images, which were generated by applying random transformations.

Optimization algorithms
To determine the optimization method more suitable to train CNNs for diagnosing melanoma, and to conclude how sensitive are these models regarding the optimization procedure used.

Transfer learning
To analyze the effectiveness of **transfer learning** paradigm for diagnosing melanoma by using pretained models, which were previously built in a different task.

Transfer learning + D.a.
To analyze the effectiveness of combining the transfer learning and data augmentation techniques for diagnosing melanoma.

Fig. 11. Phases covered in the experimental study.

tal results are analyzed, and finally some guidelines to perform melanoma diagnosis with CNN models are outlined.

5.1. Experimental design

The experimental study was divided into five phases, as shown in Fig. 11[4], encompassing each phase a different aim. First, the performance of non-adaptive and adaptive optimization methods in training CNNs for melanoma diagnosis was studied. After determining the best optimization algorithm, the second phase aimed to analyze the effectiveness of weight balancing technique for tackling the imbalance problem in melanoma datasets. Third, the power of transfer learning paradigm for diagnosing melanoma was studied. Fourth, image data augmentation techniques were applied to analyze if they can significantly improve the generalization capacity of CNNs. Finally, the last phase of the experimental study focused on determining how effective is the combination of transfer learning and data augmentation techniques in melanoma diagnosis.

In this work, twelve CNNs models (a description of each one of these models was previously provided in Section 2) were assessed on the eleven melanoma datasets previously described in Section 4. The selected models were the following: three implementations of DenseNet architecture, namely DenseNet121, DenseNet169, and DenseNet201, where the number in the name specifies the depth of the model; two versions of Inception architecture, InceptionV3 and InceptionV4; the versions ResNet50 and InceptionResNetV2 of ResNet architecture; the Xception model; two implementations of VGG architecture, namely VGG16 and VGG19; and finally, two lightweight models were selected, MobileNet, and NASNetMobile as the mobile version of NASNet architecture. This selection aimed to analyze a heterogeneous collection of CNN models presenting different features, such as different architectures (e.g. VGG, Inception, and MobileNet), different depths (e.g. different versions of DenseNet and VGG); and different complexities (in term of number of parameters to learn) ranging from big models (e.g. VGG19) to light-weight ones (e.g. MobileNet).

Experimental settings

In the first phase of the experimental study, three well-known optimization algorithms were used for training all CNN models, the well-known *Stochastic Gradient Descend* (SGD) (Goodfellow et al., 2016), *Root Mean Square Propagation* (RMSprop) (Hinton et al., 2012) and *Adaptive Moment Estimation* (ADAM) (Kingma and Ba, 2014). These three optimization algorithms has been demonstrated to work well across a wide range of deep learning architectures, serving as the basis for several others optimization methods (Goodfellow et al., 2016). SGD is a non-adaptive method that uses a common learning rate for all parameters, whereas RMSProp and ADAM are adaptive gradient descent algorithms that intend to solve the problems with sparse gradients and non-stationary settings by means of maintaining exponential moving averages of the gradient and its square. These three gradient descent algorithms iteratively compute on a small subset of training samples, named mini-batch set, instead of performing computations on the whole dataset; an epoch of the training process is completed when the entire training set is passed forward and backward through the neural network. Also, as all the studied datasets encompassed binary classification problems, the algorithms optimized a cost function that averaged a binary cross entropy along all training samples.

As for the tuning of the hyper-parameters controlling the behavior of SGD, RMSProp, and ADAM, it is noteworthy that finding the optimal set of the hyper-parameter values is a task that commonly requires expensive and arduous work due to the many possible combinations (Bergstra and Bengio, 2012). In this work, a tuning process of the hyper-parameters that regulate the behavior of these three algorithms was not carried out so the obtained results could not be conferred to an over-adjustment issue. Consequently, the basic configuration listed in Table 2 was used to train all the models for any of the experiments. It should be noted that there is no general rule-of-thumb for selecting the value of each hyper-parameter listed in Table 2 and, therefore, the choice was some-

Table 2
Configuration used in the experimental study.

Parameter	Value
Number of epochs	150
Mini-batch size	8

Learning rate (α)	ADAM={0.001, 0.01},
	RMSprop={0.001, 0.01},
	SGD=0.01

Decay rate first moment average (β_1)	ADAM=0.9, RMSprop=0.9
Decay rate second moment average (β_2)	ADAM=0.999

[4] The process chart was developed using a free template from https://www.presentationgo.com/.

thing between experience and practice. The hyper-parameters α, β_1, β_2 were set to the default values recommended in the original papers. Also, considering that the used datasets are either small or not extremely large, the batch size was set to eight samples, thus avoiding to reach local minima in early epochs, as well as avoiding frequent updates of the models' parameters that may result in noisy gradient signals and, therefore, a higher variance over training epochs. The weights of the CNNs were initialized by means of the so-called Xavier method (Glorot and Bengio, 2010), since it can achieve quicker convergence and higher accuracy on image classification tasks. Finally, to avoid the early convergence of SGD, the learning rate was reduced by a factor of 0.2 if an improvement in predictive performance was not observed during 10 epochs.

Regarding the rest of phases conducted in the experimental study, in the second phase the CNN models were forced to pay more attention to melanoma class by means of assigning weights inversely proportional to the class frequencies; therefore, although the melanoma class still has fewer observations, both classes will have the same importance in the cost function. In the third phase, transfer learning was applied by means of using pre-trained models previously built on ImageNet dataset, which contains more than 1 million of images. These models had weights that were previously learned from a different task and, it should be noted that the pre-trained models by default can predict 1000 different classes, whereas the studied problem only comprises two classes. Therefore, given a pre-trained model, the following steps were applied in order to transfer the model to melanoma diagnosis task:

- The last block of a model (block that yields the predictions, commonly composed by dense and 1D convolutional layers) was substituted by a new one with random weights.
- All the computational blocks of the model were "frozen", excepting the last block.
- The model was trained during 50 epochs (in this case, the last block was the only trainable component).
- All the blocks were defrosted and trained along the remaining 100 epochs.

In the fourth part of the study, data augmentation technique was mainly applied to tackle the imbalance problem in melanoma diagnosis by applying and combining random rotation-based, flip-based and crop-based transformations over the original images. Two types of data augmentation process were assessed, (I) performing data augmentation only in training data, and (II) performing data augmentation in both training and test data. After splitting a dataset in training and test sets, training data were balanced by creating new images until the number of melanoma images was equal to normal ones, and the generated training images were considered as independent from the original ones. On the other hand, test data were expanded by randomly augmenting each test image at least 10 times, but the generated images remained related to the original ones; i.e. images generated from a test image are not considered independent. Consequently, given an original test image X, the classes' probabilities for X and its related set of images S_X were averaged to yield the final prediction for X; so any CNN model would perform like an ensemble one, where the final probabilities for a test image was computed using a soft-voting strategy. Finally, in the fifth part of the study, the data augmentation and transfer learning techniques were combined to analyze the suitability by combining these two methods in diagnosing melanoma.

Evaluation process

Regarding the evaluation of the models, a 3-times 10-fold cross validation process was conducted on each original dataset, where the results were averaged across all fold executions. Bearing in mind the experimental design described throughout this section, a total of 39,600 CNNs were trained in this work, which can be considered as a tough and costly task.

In each fold execution, *Matthews Correlation Coefficient* (MCC) was used to measure the predictive performance of the models. MCC is widely used in Bioinformatics as a performance metric (Boughorbel et al., 2017), and it is specially designed to analyze the predictive performance on unbalanced data, even if the classes are of very different sizes. MCC gives a good summary of a confusion matrix, and it is computed as

$$MCC = \frac{t_p \times t_n - f_p \times f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}}, \tag{1}$$

where t_p, t_n, f_p, and f_n are the number of true positives, true negative, false positives, and false negatives, respectively. MCC is in range $[-1, 1]$, where 1 represents a perfect prediction, 0 indicates the model is performing similarly to random prediction, and -1 an inverse prediction. On the other hand, in those phases of the experimental study where different techniques were analyzed (phases P2-P5), the fold changes in predictive performance were assessed. Fold change (FC) describes how much the predictive performance changed after applying a different technique. It is calculated as the ratio between the predictive performance of the model applying the selected technique (e.g. transfer learning) and the performance of the baseline model (i.e. the model that did not use transfer learning). A fold change greater than 1 represents an improvement on performance, whereas a value less than 1 means the opposite case.

Non-parametric statistical tests were conducted to detect whether there were any significant difference in predictive performance. Friedman's test (Friedman, 1940) was conducted in cases where a multiple comparison was carried out (i.e. more than two models are compared), and afterward, Shaffer post-hoc test (Shaffer, 1986) was employed to perform all pairwise comparisons. On the other hand, Wilcoxon Signed-Rank test (Wilcoxon, 1945) was performed in those cases where only two individual models are compared. All hypothesis testing were conducted at 95% confidence.

Software and hardware

The experimental study was executed in two Desktop PC with Ubuntu 18.04, Intel Core i7-8700K Processor, 64 GB DDR4 RAM, and two GPUs NVIDIA Geforce GTX 1080-Ti with 11 GB DDR5 each one. All the experiments were implemented in Python v3.6, and the CNN models were developed by using Keras framework v2.2.4 (Chollet et al., 2015) as high level API, and TensorFlow v1.12 (Abadi and et al., 2015) as backend. The source code and the models trained are available at GitHub[5] and Kaggle[6], respectively.

5.2. Results

This section describes and analyzes the results obtained from the experimental study carried out. Results are summarized through boxplots, where the mean is represented by a cross and the median by a horizontal line inside the box. Due to the great amount of results collected and in order to make the paper more readable, only figures summarizing the obtained results are described[7]

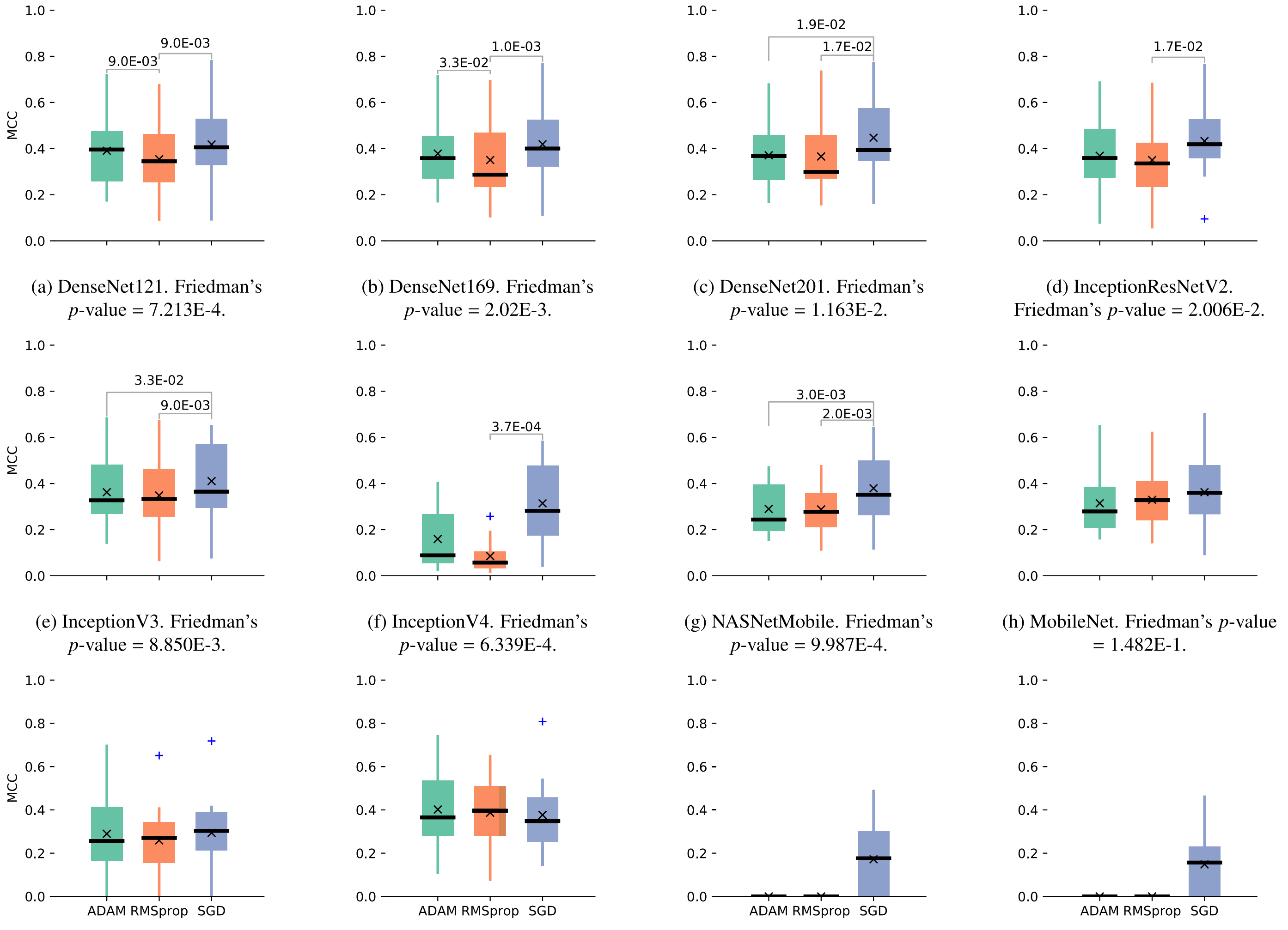
P1. Optimization algorithms for training CNNs

Fig. 12 summarizes the results obtained after training the twelve CNN models with the three selected optimization algo-

[5] https://github.com/kdis-lab/Review-CNN-Melanoma

[6] https://bit.ly/33U2bvA

[7] All the results of the experimental study can be consulted at http://www.uco.es/kdis/cnn-melanoma-review.



(a) DenseNet121. Friedman's p-value = 7.213E-4.

(b) DenseNet169. Friedman's p-value = 2.02E-3.

(c) DenseNet201. Friedman's p-value = 1.163E-2.

(d) InceptionResNetV2. Friedman's p-value = 2.006E-2.

(e) InceptionV3. Friedman's p-value = 8.850E-3.

(f) InceptionV4. Friedman's p-value = 6.339E-4.

(g) NASNetMobile. Friedman's p-value = 9.987E-4.

(h) MobileNet. Friedman's p-value = 1.482E-1.

(i) ResNet50. Friedman's p-value = 5.292E-1.

(j) Xception. Friedman's p-value = 9.624E-2.

(k) VGG16. Friedman's p-value = 8.590E-2.

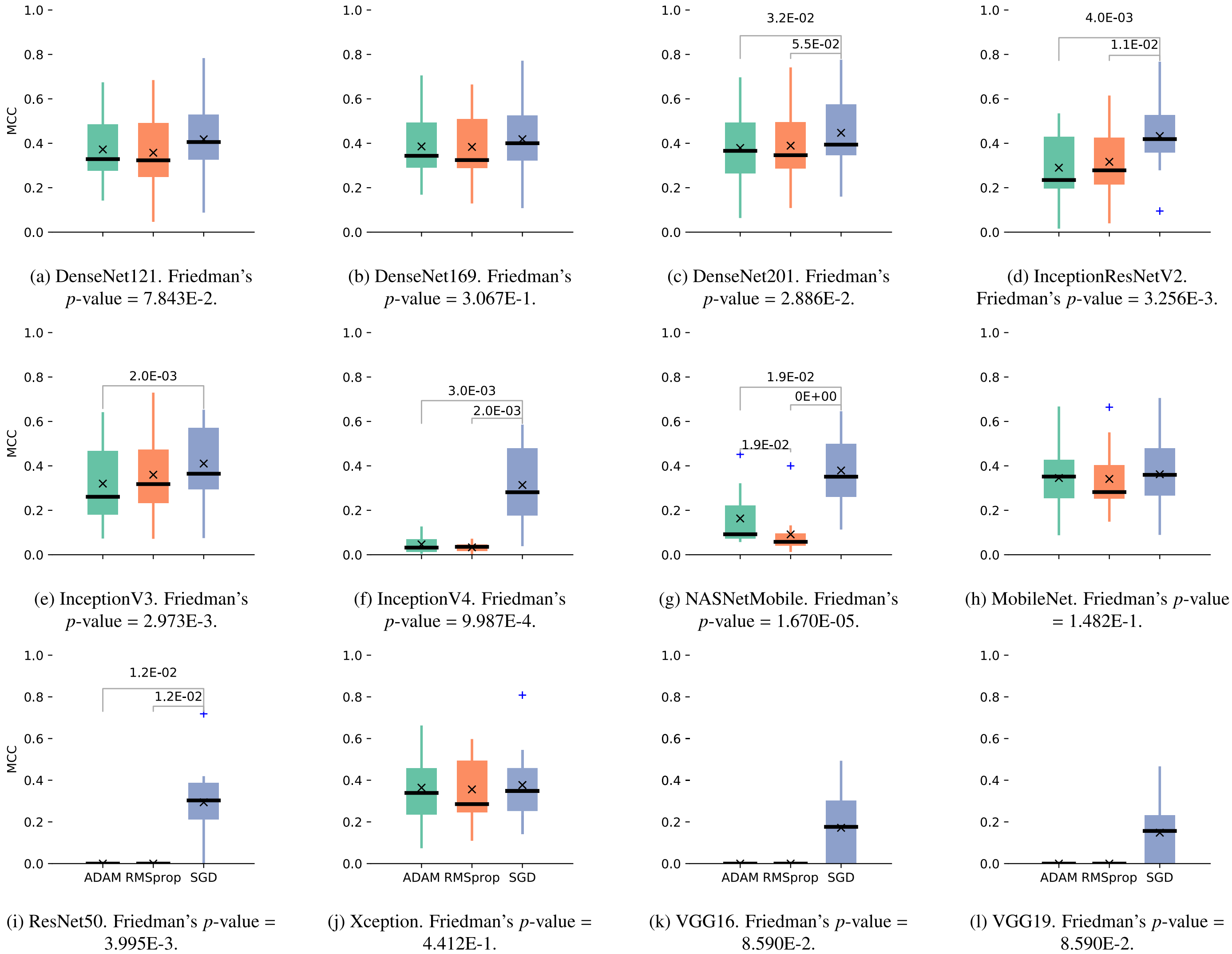
(l) VGG19. Friedman's p-value = 8.590E-2.

Fig. 12. Average MCC values on test sets by using the three optimization algorithms with the learning rates recommended by their authors. Each sub-figure summarizes the results of the Friedman's test and Shaffer's post-hoc test. Significant differences detected in pairwise comparisons are shown on top of the boxs.

rithms (SGD, ADAM and RMSprop) using the learning rates recommended by their authors. To be fair, Fig. 13 also shows the results obtained by using for ADAM and RMSprop the same learning rate of SGD. In most cases, it was observed a high variability in the results across the optimizers, and this is mainly due to the diversity and different complexities of the datasets. For example, independently of the optimizer used, most of the models achieved an acceptable predictive performance on PH2 dataset, whereas poor results were obtained on ISBI2016, ISBI2017, MSK-2, MSK-3 and MSK-4. These results matched with the previous analysis conducted on Section 4, where a more clear separation between classes was observed in PH2 dataset, whereas in ISBI and MSK datasets were observed the highest ratios between the intra-class and inter-class distances. Also, MSK-3 was the most challenge dataset, since it not only has the highest ration between intra-class and inter-class distances, but also the highest imbalance ratio, thus hampering the performance of all CNN models.

In average, SGD optimizer obtained the best average results across the models, except Xception which slightly better performed when using RMSprop (see Fig. 12). Fig. 12 shows that SGD significantly outperformed ADAM and RMSprop in three models (DenseNet201, InceptionV3, and NasNetMobile), and it also obtained better results than RMSprop in four models (DenseNet121, DenseNet169, InceptionResNetV2, and InceptionV4). Furthermore, Fig. 13 shows that SGD significantly outperformed ADAM and RMSprop in five models (DenseNet201, InceptionResNetV2, InceptionV4, NasNetMobile and ResNet50), and it also obtained better results than ADAM in InceptionV3. The least sensitive models regarding the optimizer and the learning rate used were MobileNet, Xception, VGG16, and VGG19, where no significant differences were detected between the three optimizers. Despite the proved effectiveness of the adaptive optimization methods in many other scenarios (Goodfellow et al., 2016), the overall results indicated that they generalize worse (often significantly worse) than SGD in training the CNN models for solving the tough melanoma diagnosis task. Adaptive methods had a faster initial progress on the training set, but their performance plateaued on the test set at later epochs. This result demonstrated that not always superior results can be achieved by using adaptive methods, as was previously pointed out in (Wilson et al., 2017; Keskar and Socher, 2017).

Considering SGD as the optimizer to train all the models (henceforth, the rest of the experimental study was executed using SGD optimizer), the models that achieved the best results in each dataset were the following: DenseNet121 and DenseNet169



(a) DenseNet121. Friedman's *p*-value = 7.843E-2.

(b) DenseNet169. Friedman's *p*-value = 3.067E-1.

(c) DenseNet201. Friedman's *p*-value = 2.886E-2.

(d) InceptionResNetV2. Friedman's *p*-value = 3.256E-3.

(e) InceptionV3. Friedman's *p*-value = 2.973E-3.

(f) InceptionV4. Friedman's *p*-value = 9.987E-4.

(g) NASNetMobile. Friedman's *p*-value = 1.670E-05.

(h) MobileNet. Friedman's *p*-value = 1.482E-1.

(i) ResNet50. Friedman's *p*-value = 3.995E-3.

(j) Xception. Friedman's *p*-value = 4.412E-1.

(k) VGG16. Friedman's *p*-value = 8.590E-2.

(l) VGG19. Friedman's *p*-value = 8.590E-2.

Fig. 13. Average MCC values on test sets by using the three optimization algorithms; ADAM and RMSprop used the same learning rate as SGD optimizer. Each sub-figure summarizes the results of the Friedman's test and Shaffer's post-hoc test. Significant differences detected in pairwise comparisons are shown on top of the boxs.

were the best in MSK-4 and MED-NODE datasets, respectively; DenseNet201 performed the best in HAM10000 and UDA-2; Xception obtained the best results in PH2; InceptionResNetV2 performed the best in ISBI2016, ISBI2017, MSK-1, MSK-2 and UDA-1; and finally, MobileNet was the best in MSK-3. Fig. 14 graphically represents the ranking computed by Friedman's test and the results of the multiple comparisons conducted between all the models. As can be seen, the different versions of DenseNet and InceptionResNet architectures dominated the ranking, indicating the advantages of these CNN models which are based on the addition of residual connections for each computational block, or the combination of Inception modules and residual connections. It was also observed that acceptable results can be attained by using CNN models based on depthwise separable convolution layers, like Xception and MobileNet. Furthermore, it is worth noting that lightweight models, like MobileNet and NasNetMobile, can obtain better results than InceptionV4 and Resnet50, demonstrating that not always better results are attained by using more complex models; this same conclusion can be apply for InceptionV3 and InceptionV4, where the last one had a worse performance. Finally, the two versions of VGG obtained the worst results, showing that simply stacking convolutional and pooling layers is not enough for solving the melanoma diagnosis problem.

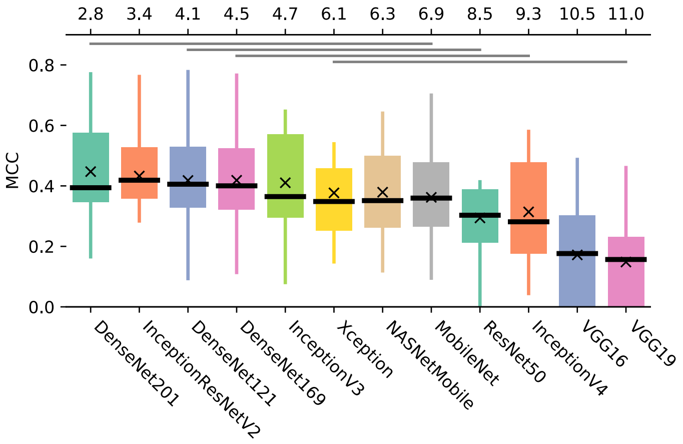
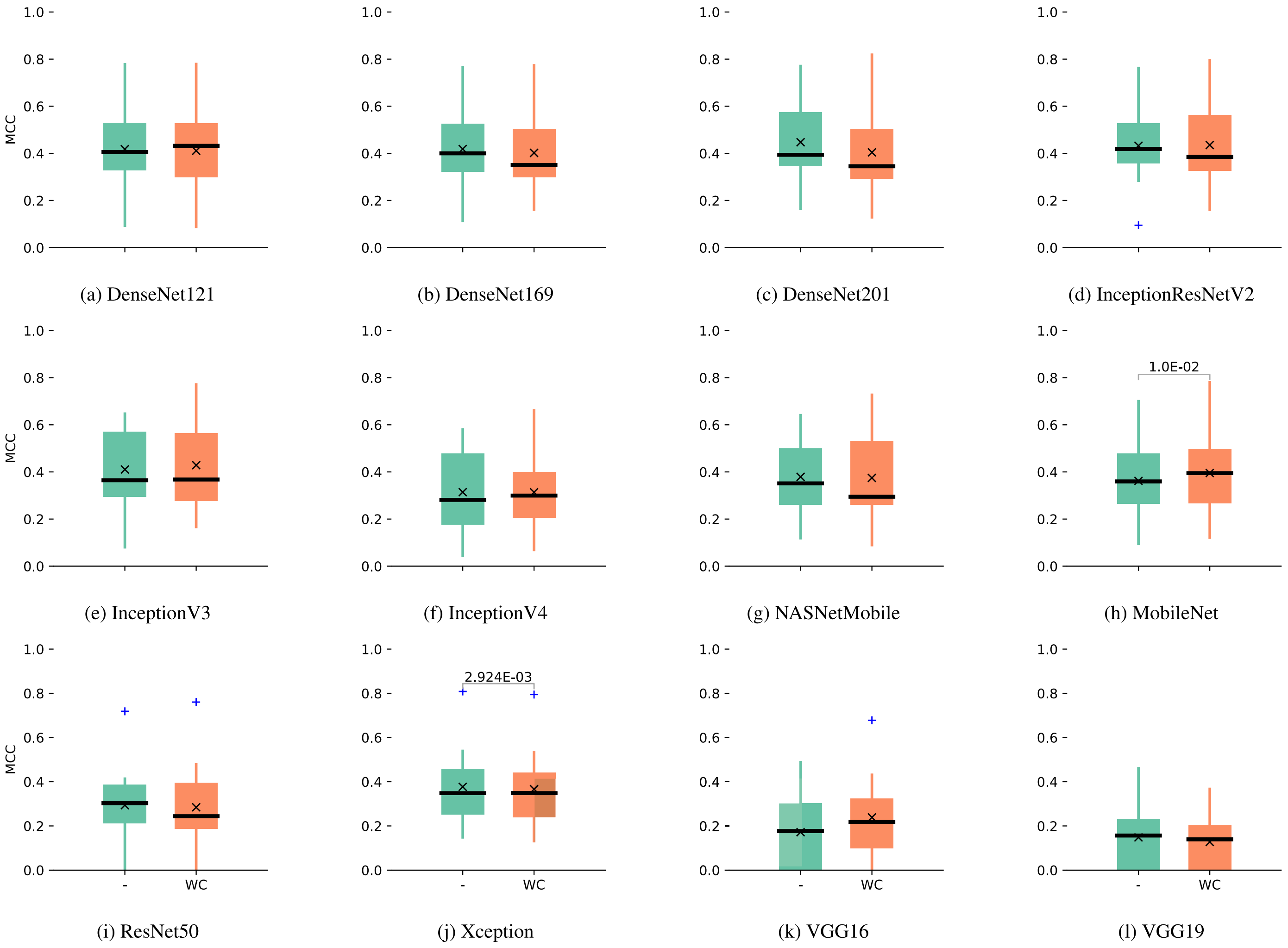


Fig. 14. All pairwise comparisons between the models; in this case, all the models used SGD optimizer. Friedman's test rejected the null hypothesis with a *p*-value equal to 5.944E-11. The models are ordered from left to right according to the ranking computed by Friedman's test (secondary x-axis located on the top of the graphic). The lines located above the boxes summarizes the significant differences encountered by the Shaffer's post-hoc test, in such a way that groups of models that are not significantly different (at α=0.05) are connected by a line.



(a) DenseNet121 (b) DenseNet169 (c) DenseNet201 (d) InceptionResNetV2

(e) InceptionV3 (f) InceptionV4 (g) NASNetMobile (h) MobileNet

(i) ResNet50 (j) Xception (k) VGG16 (l) VGG19

Fig. 15. Average MCC values on test sets by using weight balancing; "WC" represents the CNN model applied weight balancing, "-" otherwise. Significant differences detected by Wilcoxon's test are shown on top of the boxs.

P2. Weight balancing

Fig. 15 summarizes the results obtained after training the twelve CNN models but considering the weight balancing approach. The best results were obtained in PH2 and ISBI2017 datasets, where eleven and eight CNN models, respectively, improved their predictive performance regarding the baseline models; i.e. the models that did not use weight balancing. MobileNet model was the only one that, in average, significantly improved its performance by using weight balancing with respect to its baseline accuracy (see Fig. 15.h). Also, InceptionV4 and VGG16 improved their performance on eight datasets and six datasets, respectively, but no significant differences were encountered regarding their baseline performance; VGG16 obtained a performance two and three times higher than its baseline model in ISBI2016 and MSK-4, respectively. On the other hand, Xception significantly deteriorated its performance when using this approach.

Fig. 16 shows the average fold change in predictive performance for each model. The results indicated that MobileNet, InceptionV4 and VGG16 obtained the three best average fold changes, although it should be noted that the last one is characterized by the presence of extreme values. Generally speaking, the results showed that weight balancing technique did not significantly improve the performance of the CNN models, where worse average results were observed in some cases (fold changes $<$ 1), e.g. the three versions of DenseNet and Xception. These results indicated that, except for MobileNet and other isolated cases that gained in predictive performance, this cost-sensitive learning method was not an effective technique for handling the imbalance problem in melanoma diagnosis.

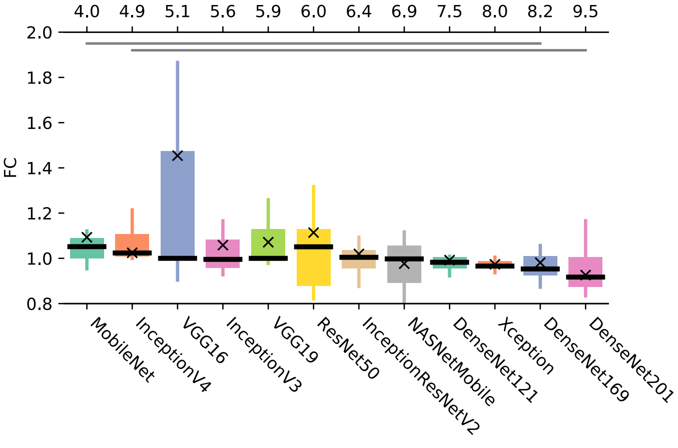
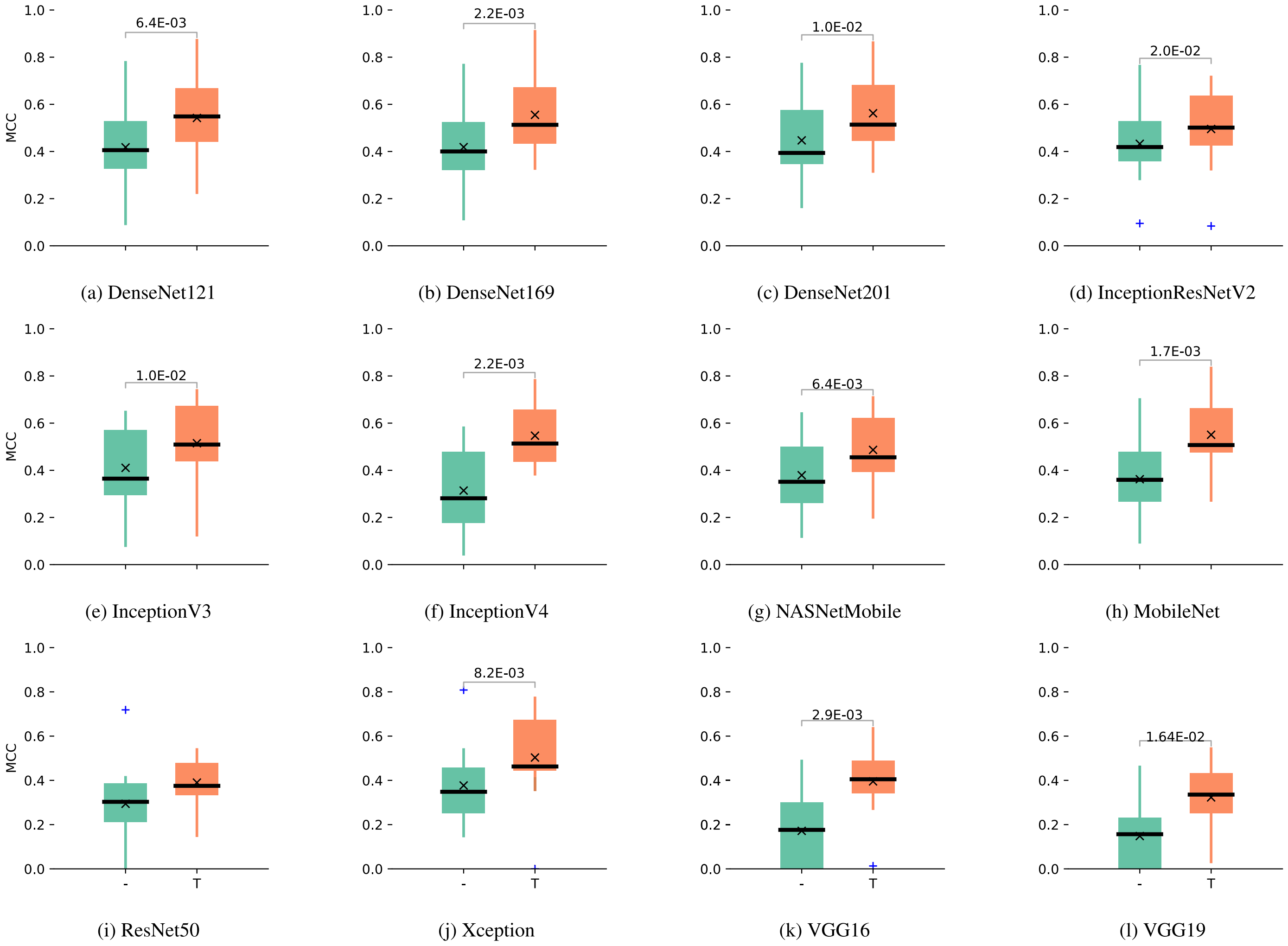


Fig. 16. Average fold changes in predictive performance by applying the weight balancing technique. Friedman's test rejected the null hypothesis with a *p*-value equal to 1.76E-2. Figure summarizes the results of the Friedman's test and Shaffer's post-hoc test.



(a) DenseNet121 (b) DenseNet169 (c) DenseNet201 (d) InceptionResNetV2

(e) InceptionV3 (f) InceptionV4 (g) NASNetMobile (h) MobileNet

(i) ResNet50 (j) Xception (k) VGG16 (l) VGG19

Fig. 17. Average MCC values on test sets by using transfer learning; "T" represents the model applying transfer learning; "-" otherwise. Significant differences detected by Wilcoxon's test are shown on top of the boxs.

P3. Transfer learning

Fig. 17 shows the results of each model after applying transfer learning. It can be observed that eleven models significantly outperformed their baseline performance. On the other hand, ResNet50 obtained in average a better performance by using transfer learning, but no significant differences were detected with respect to its baseline performance (see Fig. 17.i); ResNet50 was surpassed by its baseline model in four datasets. MobileNet was the only model that improved its performance in all the datasets, whereas VGG16, the three versions of DenseNet, and the two versions of Inception obtained better results in ten datasets. VGG16 decreased its performance by 14% in MSK-1, and DenseNet and Inception-based models had in median 19% lower performance in UDA-2. This last dataset was the most challenge, where eight model deteriorated their performance by 17% approximately; UDA-2 has the lowest Silhouette value, indicating a high overlapping level between classes. On the contrary, the easiest datasets were MSK-2, MSK-4, ISBI2016 and MED-NODE, where all the models in average increased their performance by 35% approximately. ISBI2017 was another dataset where a considerable improvement was observed by using transfer learning, e.g. DenseNet121 and InceptionV4 raised their performance from 0.17 to 0.44, and from 0.08 to 0.44 of MCC value, respectively. Summarizing, the models that achieved the best percents of improvement in each dataset by applying transfer learning were the following: InceptionV4 was the best in HAM10000 (improved its performance by 47%), MSK-1 (70%), MSK-2 (78%), MSK-3 (236%), and UDA-1 (64%) datasets; VGG16 performed the best in MED-NODE (495%), MSK-4 (344%), UDA-2 (485%) and ISBI2016 (341%) datasets; VGG19 highlighted in PH2 (76%); and finally, ResNet50 obtained the best improvement in ISBI2017 (255%).

Fig. 18 shows the average fold change in predictive performance for each model. InceptionV4, MobileNet, Xception, VGG16, and VGG19 were those most benefited from using transfer learning, where they increased in median their performance by 65%, 42%, 41%, 42%, and 56%, respectively; it should noted that the two versions of VGG architecture presented some extreme values. The results indicated that the models using depthwise separable convolution layers (MobileNet and Xception) can largely be benefited by using transfer learning. Furthermore, it is worth noting how models without a sophisticated design (the two versions of VGG) can attain better results. On the other hand, the models less benefited from using transfer learning were those based on architectures that use residual connections (models based on DenseNet and ResNet architectures). In short, despite the pre-trained models were previously built on the well-known ImageNet, which is a dataset not related with melanoma diagnosis task, the evidence showed that CNN models can outperform their baseline performance. Such an improvement in performance denoted that the 'experience" learned from ImageNet can effectively be transferred to

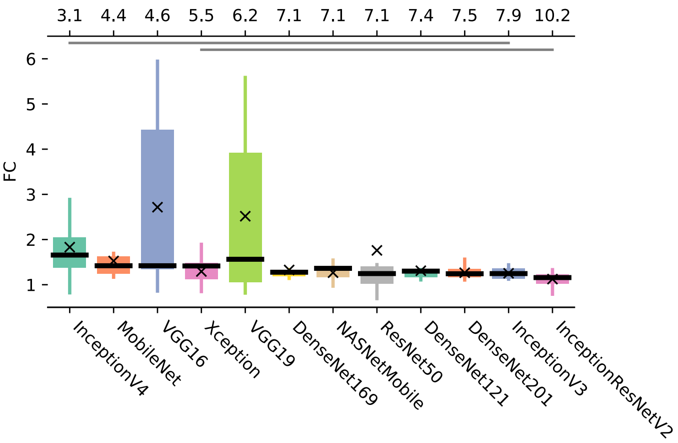


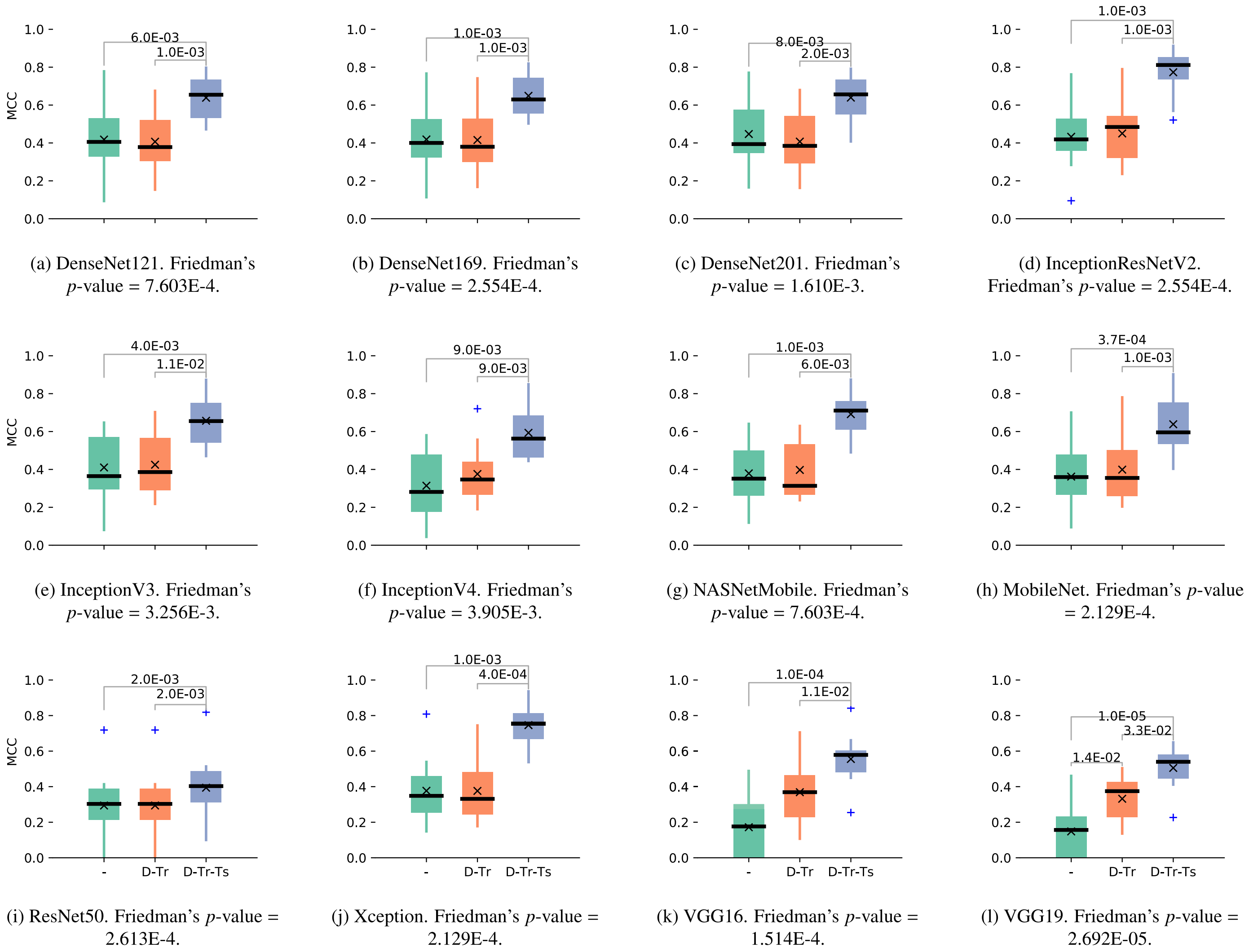
Fig. 18. Average fold changes in predictive performance by applying transfer learning. Friedman's test rejected the null hypothesis with a p-value equal to 3.863 E-4. Figure summarizes the results of the Friedman's test and Shaffer's post-hoc test.

melanoma diagnosis task by just readjusting the weights of the CNNs.

P4. Data augmentation

Fig. 19 summarizes the results attained by applying the data augmentation approach in each model (details regarding how the data augmentation process was conducted can be consulted on Section 5.1). It was observed that the models using data augmentation on both training and test sets presented a lower variability in the results, indicating a better overall improvement. Shaffer's post-hoc test confirmed that all the models significantly improved their performance when data augmentation was applied both in training and test sets. On the other hand, VGG19 model was the only one that outperformed its baseline performance by only applying data augmentation in the training set.

MSK-3 was the dataset where all the models attained a higher percent of improvement regarding their baseline performance; the models improved in median by 244%. This result is very promising because MSK-3 is the dataset with highest imbalance ratio, thus demonstrating the effectiveness of data augmentation technique for tackling the imbalance issue in melanoma datasets. Furthermore, the models in average improved their performance by



(a) DenseNet121. Friedman's p-value = 7.603E-4.

(b) DenseNet169. Friedman's p-value = 2.554E-4.

(c) DenseNet201. Friedman's p-value = 1.610E-3.

(d) InceptionResNetV2. Friedman's p-value = 2.554E-4.

(e) InceptionV3. Friedman's p-value = 3.256E-3.

(f) InceptionV4. Friedman's p-value = 3.905E-3.

(g) NASNetMobile. Friedman's p-value = 7.603E-4.

(h) MobileNet. Friedman's p-value = 2.129E-4.

(i) ResNet50. Friedman's p-value = 2.613E-4.

(j) Xception. Friedman's p-value = 2.129E-4.

(k) VGG16. Friedman's p-value = 1.514E-4.

(l) VGG19. Friedman's p-value = 2.692E-05.

Fig. 19. Average MCC values on test sets by using data augmentation; "-" represents the baseline model; "D-Tr" represents the results applying data augmentation only in training set; "D-Tr-Ts" represents the results applying data augmentation in both training and test sets. Each sub-figure summarizes the results of the Friedman's test and Shaffer's post-hoc test.

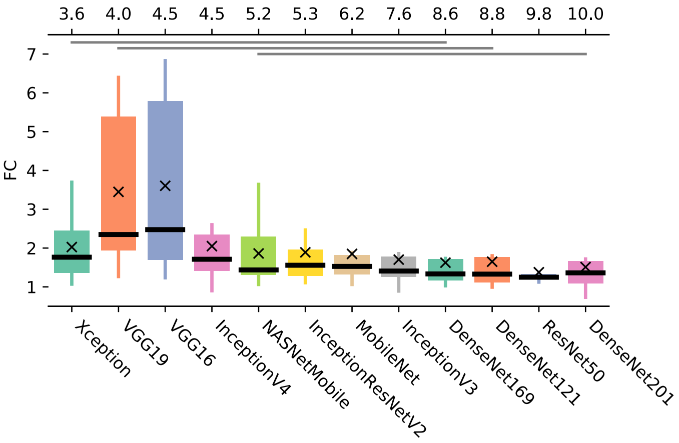


Fig. 20. Average fold changes in predictive performance by applying data augmentation. Friedman's test rejected the null hypothesis with a *p*-value equal to 4.698E-07. Figure summarizes the results of the Friedman's test and Shaffer's post-hoc test.

48%, 159%, and 83% in HAM10000, ISBI2017, and ISBI2016, respectively, which are the three other datasets with highest imbalance ratio. On the other hand, as expected the lowest improvement in performance was observed in MED-NODE, since it is a small dataset and has the lowest imbalance ratio and, therefore, the application of data augmentation approach was not so effective in this case. All in all, the models that achieved the best percents of improvement in each dataset by applying data augmentation on training and test sets were the following: Xception was the best in HAM10000 (increase its baseline performance by 76%) and MSK-2 (157%) datasets; InceptionV4 obtained the best results in MSK-2 (71%); InceptionResNetV2 performed the best in MSK-3 (371%); VGG16 was the best in MED-NODE (579%), MSK-4 (583%), and UDA-2 (450%) datasets; VGG19 highlighted in PH2 (126%), UDA-1 (74%), and ISBI2016 (511%) datasets; and finally, MobileNet was the best in ISBI2017 (333%).

Fig. 20 shows the average fold change in predictive performance for each model by applying data augmentation both on training and test sets. It can be observed that Xception was the model most benefited by using this technique (increased in median its performance by 76%), followed by VGG19 (135%) and VGG16 (147%); it should noted that the high values observed on the two versions of VGG architecture are due to some extreme values. Also, the least benefited models from using data augmentation were those based on architectures that use residual connections (models based on DenseNet and ResNet architectures); these results are similar to the ones obtained in the phase applying transfer learning. In short, the results indicated that data augmentation is an effective approach to tackle the imbalance issue in melanoma diagnosis task, and also that an alluring performance can be achieve by not only augmenting data on the training sets but also on the test sets. Performing data augmentation technique on test sets provokes that CNN models performed similar to ensemble models, obtaining a final prediction by aggregating the partial predictions yielded from an original image and all its random transformations and, therefore, providing a higher robustness to the inter- and intra-class variability present in melanoma images.

P5. Combining transfer learning and data augmentation

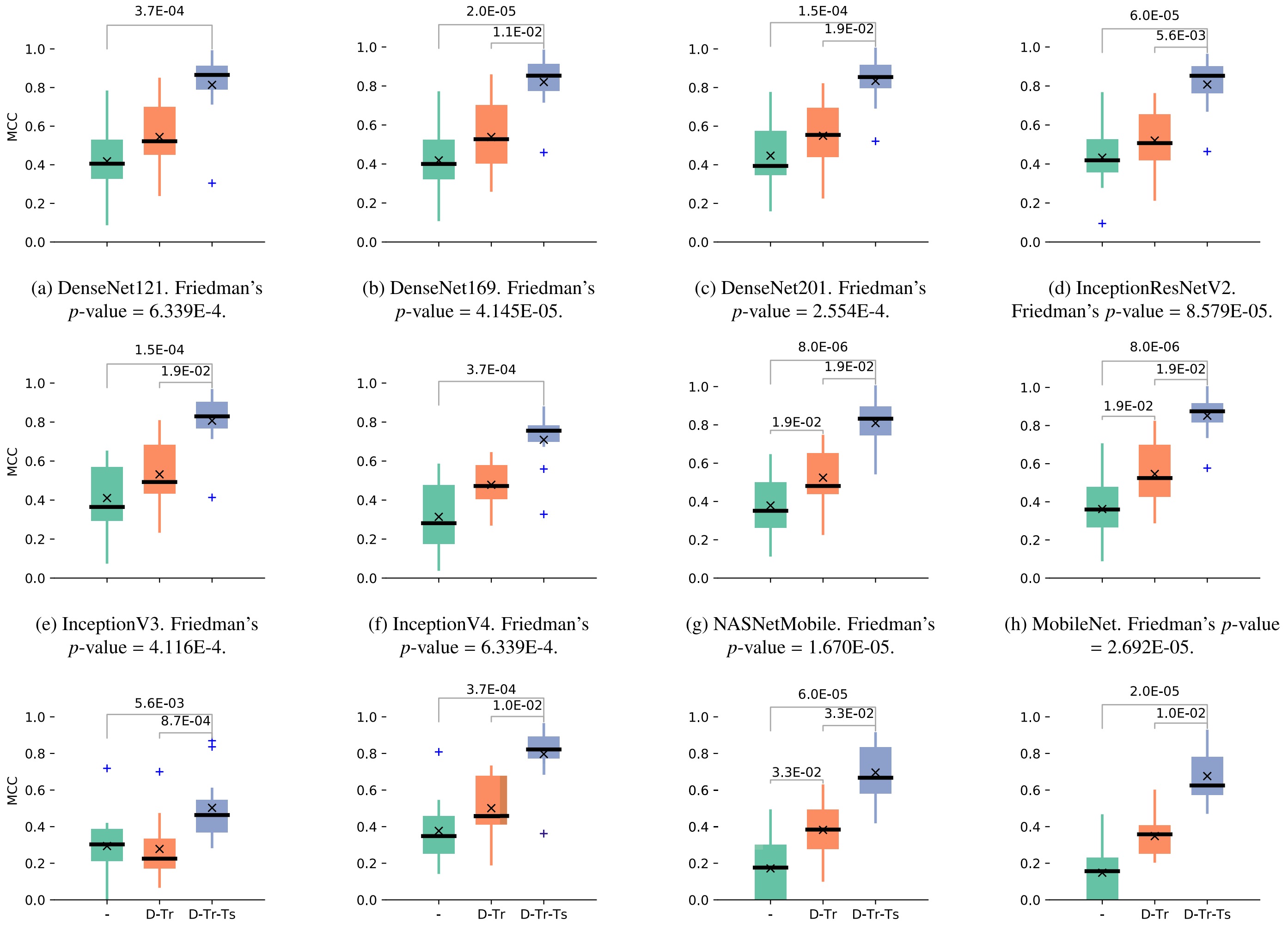
Fig. 21 shows the results of combining transfer learning and data augmentation. Similar to the results pointed out in the previous section, all the models applying data augmentation both on training and test sets significantly outperformed their baseline performance. However, in this case, it was also observed that there were two models where no significant differences were encountered when applying data augmentation only on training set or on both sets (DenseNet201 and InceptionV4). Also, three models showed a significant improvement regarding their baseline versions by only performing data augmentation on training sets (NasNetMobile, MobileNet and VGG16).

MSK-3 was the dataset where all the models in average obtained the highest improvement by applying transfer learning and data augmentation both on training and test sets (424%), followed by ISBI2017 (303%), MSK-2 (232%), ISBI2016 (227%), MSK-4 (219%) and UDA-1 (78%). All these datasets have the highest imbalance ratios, thus indicating that the combination of these two technique is very suitable for tackling the imbalance problem in melanoma diagnosis. On the other hand, the lowest improvement was observed in UDA-2 (only by 2%), and this result was expected because this dataset is the smallest one and does not have a moderate imbalance problem. In short, the models that achieved the best percents of improvement in each dataset by applying transfer learning and data augmentation on training and test sets were the following: VGG19 was the best in MSK-3 (increased its performance by 911%), PH2 (236%), UDA-2 (477%) and ISBI2016 (625%) datasets; InceptionV4 obtained the best results in MSK-1 (91%), MSK-2 (176%) and UDA-1 (43%) datasets; VGG16 was the best in MED-NODE (611%) and MSK-4 (906%) datasets; and finally, ResNet50 obtained the highest improvement in HAM10000 (90%) and ISBI2017 (414%) datasets (See Fig. 22).

Fig. 22 shows the average fold change in predictive performance for each model. It can observed that the most benefited models were the two versions of VGG (in average, approximately increased their performance by 220%), and MobileNet (91%); it should noted that the two versions of VGG architecture presented some extreme values. It is worth noting how models without a sophisticated design (the two versions of VGG) can attain better results. Also, the least benefited models from combining the two techniques were those based on architectures that use residual connections (models based on DenseNet and ResNet architectures); these results are similar to the ones obtained in the phases where transfer learning and data augmentation were applied individually.

5.3. Summary and suggestions

Fig. 23 shows the average performance attained by the models on dermoscopic and non-dermoscopic datasets. Ten datasets comprise dermoscopic images and, therefore, the results obtained on these datasets were averaged, whereas only MED-NODE dataset comprises non-dermoscopic images. In the three first phases of the experimental study, it was observed that the overall performance of the CNNs was higher on non-dermoscopic images, except for ResNet50, VGG16 and VGG19. It was also observed the application of transfer learning was very helpful to edge up the prediction level of the models on non-dermoscopic images, indicating that the patterns previously learned on ImageNet had a very positive impact; ImageNet and MED-NODE datasets are composed by images shot with common digital cameras. However, the performance attained was inferior when data augmentation was applied on non-dermoscopic images. This revealed that if only a data augmentation was performed over non-dermoscopic melanoma images, a significant improvement in the CNNs' performance could not be attained, and this is due to the quality of the images shot with common cameras, which may have several issues, such as poor lighting, grainy, blurry, pixelated or not centered images. Finally, the performance over non-dermoscopic images again edged up when transfer learning and data augmentation were applied together. As for the results on dermoscopic images, data augmentation process showed to be a very effective method to improve the prediction levels. Finally, the results showed the combination of transfer learning and data augmentation can be very benefi-



(a) DenseNet121. Friedman's *p*-value = 6.339E-4.

(b) DenseNet169. Friedman's *p*-value = 4.145E-05.

(c) DenseNet201. Friedman's *p*-value = 2.554E-4.

(d) InceptionResNetV2. Friedman's *p*-value = 8.579E-05.

(e) InceptionV3. Friedman's *p*-value = 4.116E-4.

(f) InceptionV4. Friedman's *p*-value = 6.339E-4.

(g) NASNetMobile. Friedman's *p*-value = 1.670E-05.

(h) MobileNet. Friedman's *p*-value = 2.692E-05.

(i) ResNet50. Friedman's *p*-value = 7.603E-4.

(j) Xception. Friedman's *p*-value = 4.826E-4.

(k) VGG16. Friedman's *p*-value = 1.127E-4.

(l) VGG19. Friedman's *p*-value = 4.145E-05.

Fig. 21. Average MCC values on test sets by using transfer learning and data augmentation; "-" represents the baseline model; "D-Tr" represents the results applying transfer learning and data augmentation only in training set; "D-Tr-Ts" represents the results applying transfer learning and data augmentation in both training and test sets. Each sub-figure summarizes the results of the Friedman's test and Shaffer's post-hoc test.

cial for building better models either on dermoscopic and non-demorscopic images.

Fig. 24 shows a summary of the performance attained by the models when applying each one of the techniques considered in the experimental study. According to the twelve Friedman's rankings (one Friedman's ranking for each model), it was confirmed that all the models attained the best performance when combining transfer learning with data augmentation. Also, data augmentation and transfer learning (applied individually), were the second and third best techniques, respectively; except in DenseNet201, that slightly deteriorated its performance when applying data augmentation on HAM10000, PH2 and UDA-2 datasets. In average, the models applying transfer learning or data augmentation, or the combination of these two, obtained a better predictive performance than when using weight balancing and their baseline versions. Weight balancing, on the other hand, did not demonstrate to improve significantly the learning of CNNs models in melanoma diagnosis. ResNet50 was the model less benefited from applying the four techniques studied.

Fig. 25 shows a comparison between all the models when applying transfer learning and data augmentation at a time. Although Shaffer's test did not detected significant differences in several cases, it is very impressive to see MobileNet located at the first position of the Friedman's ranking, showing the effectiveness of this light-weight model for diagnosing melanoma. DenseNet-based models and InceptionResnetV2 also obtained very good results, showing the benefits of using either residual connections or the combination of this type of connection with Inception modules. It was also interesting to see the light-weight NASNetMobile reached better results than other models more complex such as Inception-based models, Xception and ResNet50.

All in all, the CNN models improved their performance compared to their baseline versions (i.e. those models that did not applied any extra method for learning) in tough datasets by means of applying techniques such as transfer learning and data augmentation. For example, it was observed that the best improvement by far was attained in MSK-3 and ISBI-2017 datasets. In average, the models applying transfer learning increased their baseline performance about 237% and 231% in ISBI-2017 and MSK-3, respectively. The improvement was even higher when applying data augmentation, with nearly 589% and 414% of improvement in MSK-3 and ISBI-2017, respectively. Moreover, the best results were achieved when both techniques were applied simultaneously, obtaining 907% and 484% of improvement in MSK-3 and ISBI-2017,

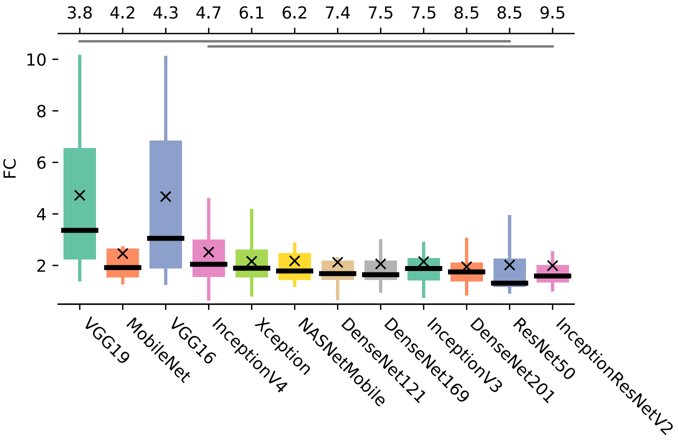
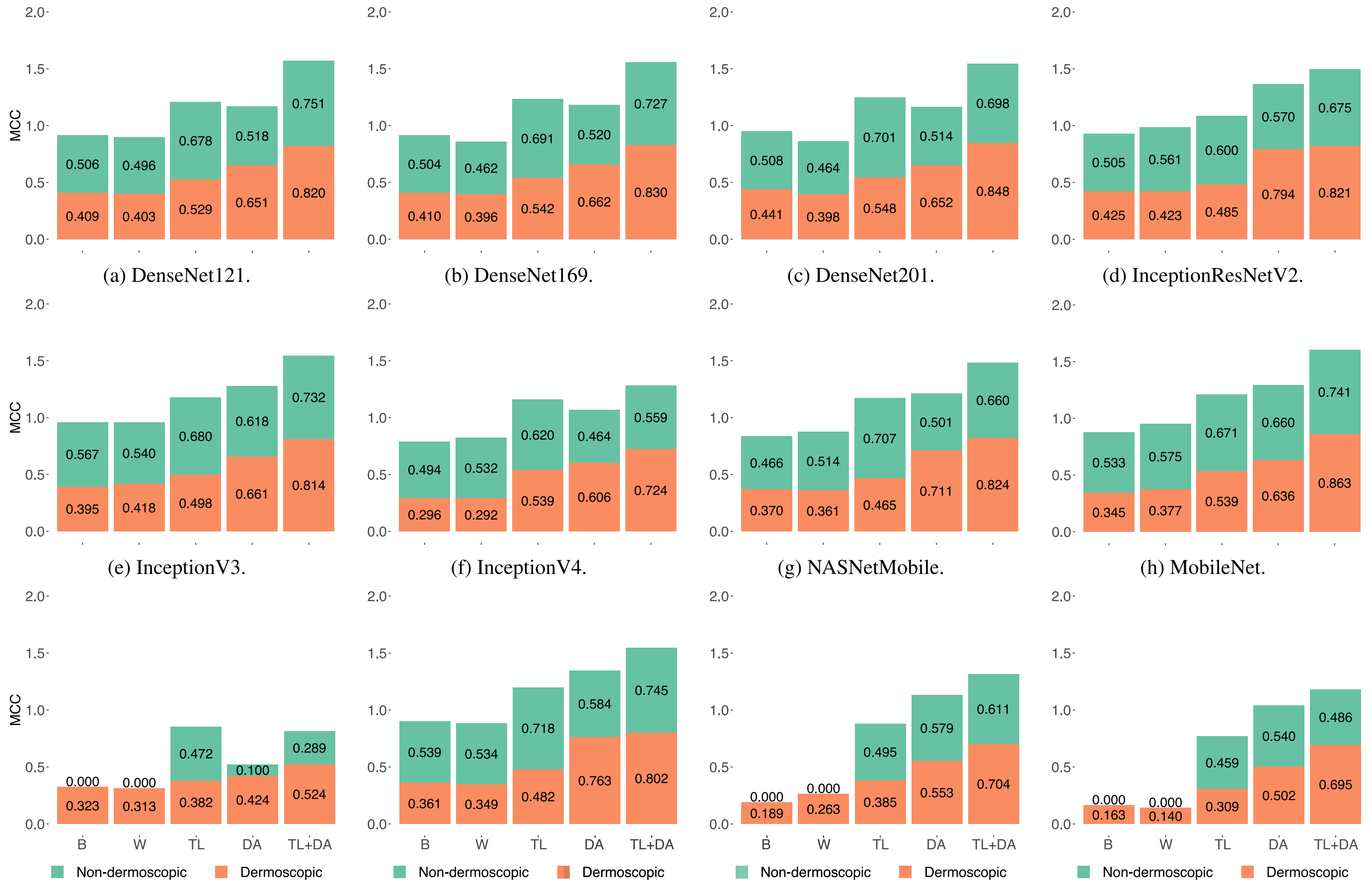


Fig. 22. Average fold changes in predictive performance by applying transfer learning and data augmentation techniques. Friedman's test rejected the null hypothesis with a *p*-value equal to 3.551E-4. Figure summarizes the results of the Friedman's test and Shaffer's post-hoc test.

respectively. It should be noted that MSK-3 was the dataset with the highest imbalance factor (approximately 11x) and, therefore, more images were generated to balance the data, so demonstrating the suitability of the data augmentation process for balancing data and improving the CNN models. Regarding ISBI-2017 dataset, it is the third most unbalanced dataset and one of the datasets with lowest Silhouette score. Accordingly, the results showed that techniques such as data augmentation and transfer learning can significantly improve the CNN models' performance in datasets with a moderate or high imbalance problem and a high overlapping level between classes. On the other hand, in average the lowest improvement was attained in UDA-2 and MED-NODE datasets. Transfer learning outperformed the baseline performance by only 86% and 132% in UDA-2 and MED-NODE, respectively, whereas data augmentation improved about 104% and 107% the baseline performance. Moreover, when transfer learning and data augmentation techniques were applied together an improvement of only 98% and 136% in UDA-2 and MED-NODE, respectively, were achieved. These results are related with the fact that UDA-2 and MED-NODE were the smallest and more balanced datasets used in the study, and it is well known that the larger datasets the better the performance that may achieved by deep learning models.

In addition, the required training time (during 150 epochs) and used GPU memory were analyzed to measure the computational complexity of each CNN model. Fig. 26 shows the average training time for each CNN model on the eleven datasets. As expected, MobileNet was the fastest CNN model, with an average training time of approximately 26 minutes. However, unexpectedly NASNetMobile obtained the second highest training time, indicating that although it is a lightweight model the internal reinforcement learning method used to search an effective building block is compu-

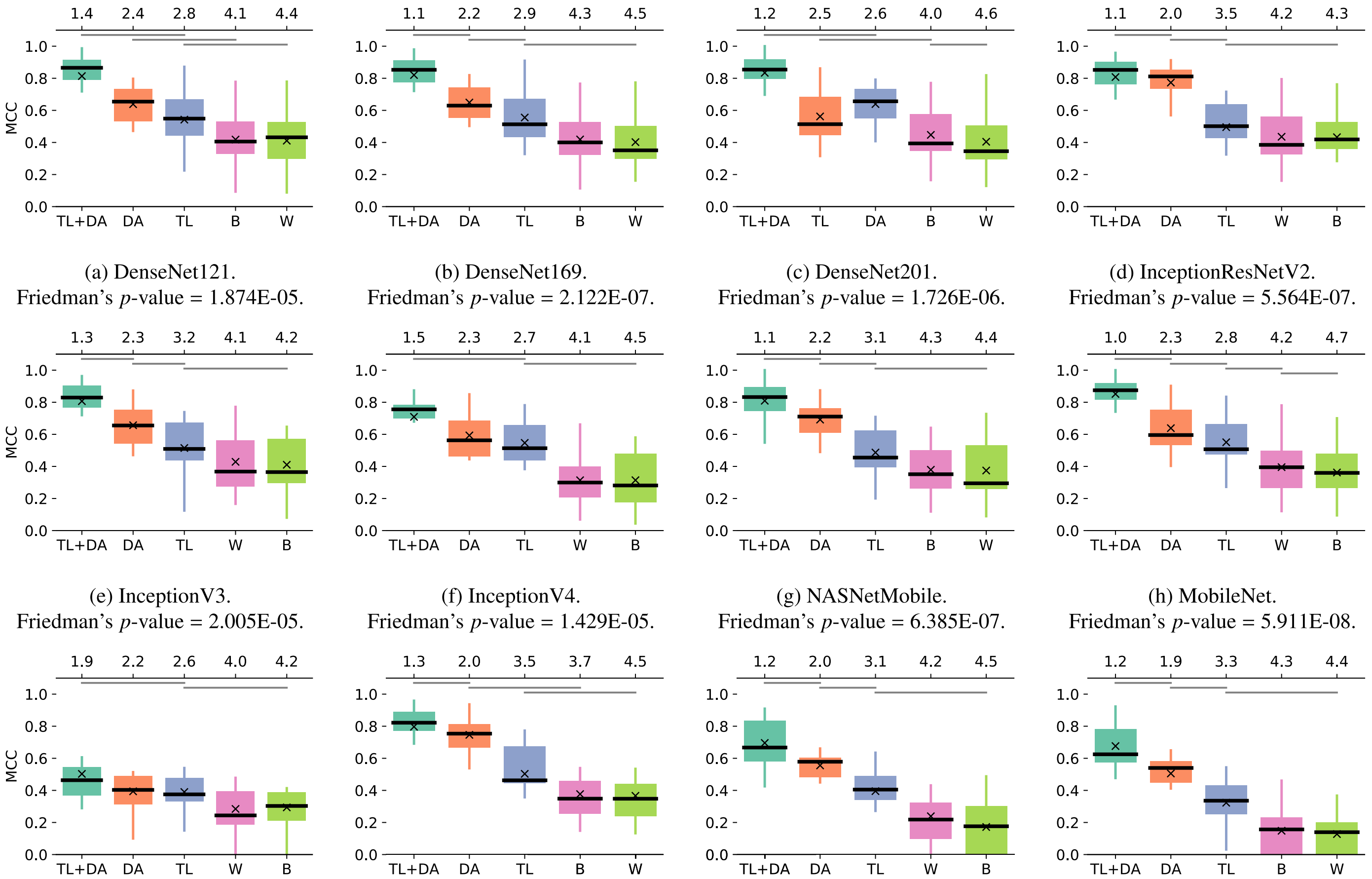


(a) DenseNet121. (b) DenseNet169. (c) DenseNet201. (d) InceptionResNetV2.

(e) InceptionV3. (f) InceptionV4. (g) NASNetMobile. (h) MobileNet.

(i) ResNet50. (j) Xception. (k) VGG16. (l) VGG19.

Fig. 23. Average MCC values on test sets considering datasets of dermoscopic (orange bars) and non-dermoscopic (green bars) images; "B" represents the baseline model; "W" represents the results applying weight balancing; "TL" represents the results applying transfer learning; "DA" represents the results performing data augmentation on training and test phases; "TL+DA" shows the results combining transfer learning and data augmentation. Each sub-figure summarizes the results of each CNN model along the five phases studied in the experimental study. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) DenseNet121.
Friedman's *p*-value = 1.874E-05.

(b) DenseNet169.
Friedman's *p*-value = 2.122E-07.

(c) DenseNet201.
Friedman's *p*-value = 1.726E-06.

(d) InceptionResNetV2.
Friedman's *p*-value = 5.564E-07.

(e) InceptionV3.
Friedman's *p*-value = 2.005E-05.

(f) InceptionV4.
Friedman's *p*-value = 1.429E-05.

(g) NASNetMobile.
Friedman's *p*-value = 6.385E-07.

(h) MobileNet.
Friedman's *p*-value = 5.911E-08.

(i) ResNet50.
Friedman's *p*-value = 4.560E-4.

(j) Xception.
Friedman's *p*-value = 2.693E-06.

(k) VGG16.
Friedman's *p*-value = 3.651E-07.

(l) VGG19.
Friedman's *p*-value = 3.811E-07.

Fig. 24. Average MCC values on test sets. "B" means a baseline performance (i.e. none technique was applied). "W", "TL", "DA" and "TL+DA" represent the model applying weight balancing, transfer learning, data augmentation, and combining transfer learning and data augmentation, respectively; data augmentation was applied both in training and test sets. Each sub-figure summarizes the results of the Friedman's test and Shaffer's post-hoc test.

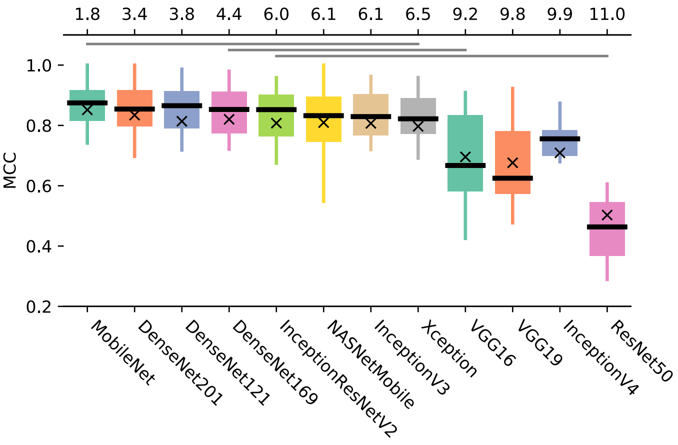


Fig. 25. Average MCC values of each model by combining transfer learning and data augmentation. Friedman's test rejected the null hypothesis with a *p*-value equal to 1.751E-12. Figure summarizes the results of the Friedman's test and Shaffer's post-hoc test.

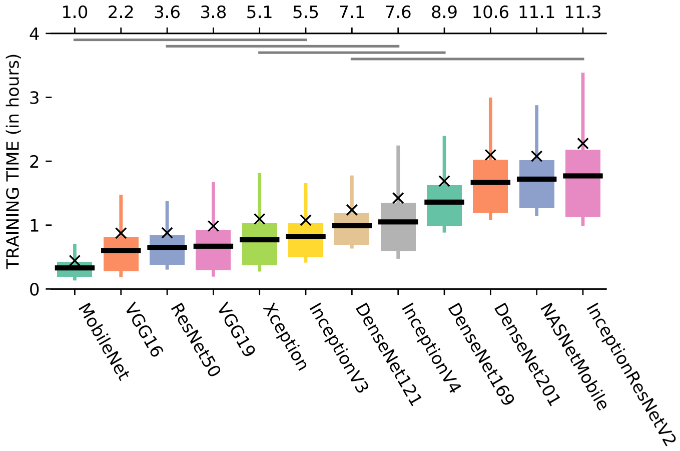


Fig. 26. Average run time (in hours) for training (during 150 epochs) each CNN model. Each box-plot summarizes the average training times in the eleven datasets. Friedman's test rejected the null hypothesis with a *p*-value less than $<$ 2.2E-16. Figure summarizes the results of the Friedman's test and Shaffer's post-hoc test.

tationally complex. On the other hand, InceptionResNetV2 had the worst training time, with an average time of about 137 minutes and being five times slower compared to MobileNet. It was interesting to see that MobileNet and VGG16 were significantly faster compared to InceptionV4, NASNetMobile, InceptionResNetV2 and all versions of DenseNet. Furthermore, the results indicated that the larger the datasets the more training time required to train the models with more complex architectures; e.g. InceptionRestNet model. On the other hand, Table 3 shows the minimum amount

Table 3
GPU memory used by the CNN models. The models are arranged in increasing order by GPU memory usage.

CNN	Parameters (millions)	GPU memory (megabytes)
MobileNet	4	671
NASNetMobile	5	703
ResNet50	25	873
DenseNet121	8	933
DenseNet169	14	1015
DenseNet201	20	1043
InceptionV3	23	1085
InceptionV4	41	1241
Xception	22	1271
InceptionResNetV2	55	1437
VGG16	138	2051
VGG19	143	2099

of GPU memory required by each CNN model. As expected, MobileNet and NASNetMobile were the light-weight models which required the least amount of GPU memory, whereas VGG-based models consumed the highest amount of GPU memory. Bear in mind that the amount of GPU memory needed is directly related with the number of trainable parameters. However, despite some CNN models required several gigabytes of GPU memory (e.g. Inception and DenseNet based architectures), it should be noted that the memory used was much lower than the available resources that can be encountered in modern and powerful GPU cards.

The above results showed that several CNN models are suitable for melanoma diagnosis. However, an important task still remains for the correct application and deployment of CNNs in real-world environments, that is the explanation of the individual predictions made by the models. This task is crucial for assessing trust on the predictions yielded and, therefore, for an effective interaction of the biomedical experts with machine learning systems. In this regard, several methods can be used to explain the predictions of black-box models (Guidotti et al., 2019), e.g. by using Accumulated Local Effects (Apley, 2016), Feature Interaction (Greenwell et al., 2018), Permutation Feature Importance (Fisher et al., 2018), Local Surrogate Models (Ribeiro et al., 2016), Counterfactual Explanations (Wachter et al., 2017), Individual Conditional Expectation (Goldstein et al., 2015), Influential Instances (Koh and Liang, 2017), and Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017). For example, Fig. 27 shows how five models learned the features (pixels or regions of the images) that allow to correctly predict whether a sample is melanoma or not. The models adjusted their weights along the training epochs, and finally they focused on the most important features to give a prediction with a high confidence level. Now, the experts not only can analyze the probability that a test sample is melanoma or not, but also assess the rationale followed by the model to give such a conclusion. In this case, it was observed that in the first four images the models focused on the lesion areas, whereas in the last input image the model payed more attention on the non-lesion area; red pixels increase the probability that a sample is melanoma, whereas the blue pixels the opposite.

To sum up, throughout the extensive experimental study conducted in this work it was demonstrated the automatic diagnosis of melanoma is a tough task. However, it was shown that such a difficulty can be reduced by applying different techniques when training the CNN models. Accordingly, the following suggestions can be drawn from this study:

- We encourage the use of evaluation metrics that are able to capture the real performance of classifiers when solving the melanoma diagnosis problem, avoiding as possible the bias induced by majority classes. In this work, MCC was used as evaluation metric since it can effectively analyze the predictive performance on unbalanced data, even if the classes are of very different sizes. However, MCC and the well-known Area Under the Receiver Operating Characteristics (AUC) consider all classes equally important. In this regard, some other interesting measures can also be used such Cohen's kappa rate (Ben-David, 2008) or balanced accuracy (Brodersen et al., 2010) to take into account the imbalance issue. Finally, precision and recall measures can be used to pay more attention to a specific class.
- Some CNN models demonstrated to be less sensitive to the selection of the optimization algorithm used for learning their weights; e.g. MobileNet, ResNet, Xception and VGG. However, the overall results indicated that better performance were attained in average when using SGD optimizer. Therefore, we recommend to first consider non-adaptive optimization methods for training CNN models, but always reducing the learning rate to avoid an early convergence.
- Weight balancing technique should be used with caution, because there is not evidence that this cost-sensitive method can induce a better melanoma diagnosis; even in some cases, the performance could be affected.
- Transfer learning and data augmentation are very effective technique to improve the melanoma diagnosis, allowing the construction of more robust CNN models. Performing data augmentation technique on both training and test sets can provide a higher robustness to the inter- and intra-class variability present in melanoma images, and also it is an excellent approach to cope with the imbalance issue. On the other hand, transfer learning allows to use powerful patterns (weights) and subsequently readjust them to construct more discriminative ones, thus reducing the cost of constructing accurate classifiers. Finally, we recommend the combination of these two technique as a powerful tool for a better melanoma diagnosis.
- HAM10000 and PH2 datasets can serve to extract useful patterns (weights) that can subsequently be transferred for tackling more challenging datasets such as those belonging to MSK and UDA collections. The former two datasets have the two highest Silhouette scores and, therefore, more discriminative patterns can be extracted from them.
- We recommend the use of CNN architectures based on residual connections for a better melanoma diagnosis, yet specifically those ones based on InceptionResNet or DenseNet. The results showed the effectiveness that can be attained when combining residual connections with Inception modules (InceptionResNet architecture), or when each building block receives additional inputs form all preceding blocks (DenseNet architecture). Inception architecture enables the extraction of different features by using different kernels at the same time, whereas the residual connections allow deeper blocks can work with non-preprocessed information, which can lead to a mitigation of the vanishing-gradient problem, a propagation and reuse of features, and a reduction of the number of parameters.
- The use of MobileNet architecture for melanoma diagnosis is strongly encouraged; this model achieved a good performance throughout the conducted study. Depthwise separable convolutions enable the reduction of the model size and the complexity for training. By this way, effective light-weight models embedded into mobile devices could be developed.
- Finally, four interesting research lines on which the researchers should pay more attention for a better training of CNN models are ensemble learning, multi-view learning, generative adversarial networks, and multi-task learning. Also, the development of agnostic tools for explaining the final predictions yielded by CNNs is another field of study of extreme importance in order to attain a better application, exploitation and acceptance of this type of machine learning model in biomedicine.

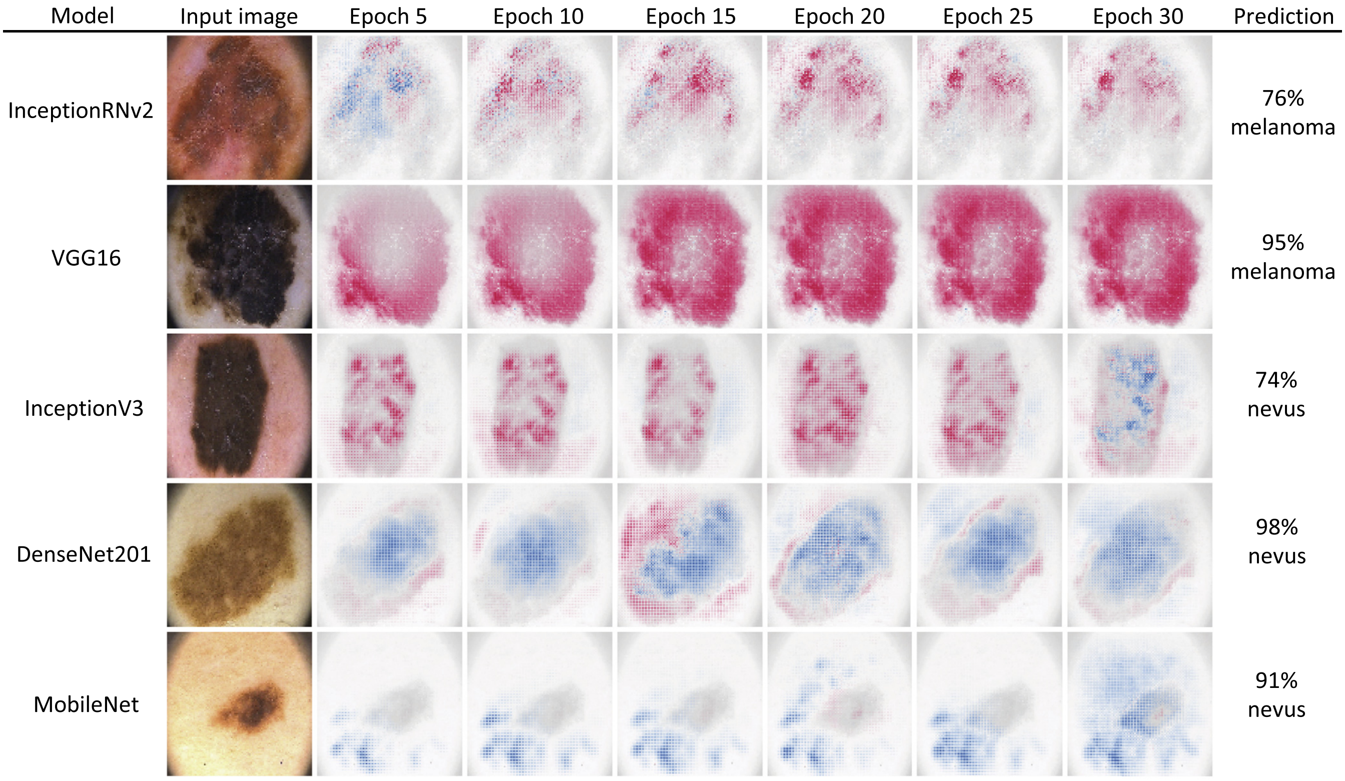


Fig. 27. Example that shows how the models learned the most important features along the training epochs until they made a final prediction. The five input images were selected from PH2 dataset, and the models correctly predicted their true labels (melanoma, melanoma, normal nevus, normal nevus and normal nevus). In this case, SHAP method was used to visualize and explain the predictions given by the models. SHAP is a model-agnostic interpretation tool that computes Shapley values (Shapley, 1953), thus having a solid theoretical foundation in game theory, and can be applied to any supervised machine learning model. A shapley value denotes how much a feature value contributed to the prediction of an example compared to the average prediction for the dataset. More details about SHAP method can be consulted in (Lundberg and Lee, 2017).

Based on the above suggestions, it is recommended to design novel CNN architectures for diagnosing melanoma by means of combining the following advanced techniques: multi-view learning, ensemble learning, data augmentation and transfer learning. It would be interesting to see a CNN model that generates multiple views from each original image by means of applying independent and specialized transformations, and its main goal would be to learn one function that models each view and jointly optimize all the functions to improve the generalization performance. For constructing the different views of an image, the model could internally apply a data augmentation process guided by a heuristic search method (e.g. a genetic algorithm) that would allow to find and learn the best sets of specific transformations to be applied on each original image. Accordingly, each generated view would represent a different feature space, thus reducing the large number of representations that would be generated if all possible transformations were simultaneously applied over the original images, and making the model more transformation-invariant. Furthermore, it is important that the CNN model would implicitly follow an ensemble approach by means of learning several auxiliary classifiers (weak classifiers) over each generated view. Consequently, the final prediction for each original image could be yielded by aggregating information that comes from each auxiliary classifier, thus allowing to attain a more stable learning and better convergence, as well as a better regularization effect. Finally, it would be important to apply transfer learning by means of extracting computational blocks (i.e. a serie of convolutional layers) from CNN models that were pretrained on different large scale datasets. Hence, heterogeneous pretrained computational blocks (e.g. computational blocks based on depthwise separable convolutions, residual connections and inception modules) would be used to build the auxiliary classifiers which learn over the generated views.

6. Conclusions

In this work, an extensive experimental review was carried out, aiming at assessing the suitability of CNN models in melanoma diagnosis. The experimental study assessed twelve CNN models on eleven public image datasets. First, several features were measured on the collection of datasets, and they confirmed the high complexity of the melanoma diagnosis task, showing that the underlying classification problems are commonly characterized by a high inter- and intra-class variability. Second, the sensitivity of CNN models regarding the optimization method was analyzed, demonstrating that not always superior performance can be achieved by using adaptive methods. Third, several techniques were assessed to determine the impact on the performance of CNN models, and the results showed that a great improvement can be attained when transfer learning and data augmentation techniques are applied together.

Consequently, through this experimental review, it has been confirmed the usefulness, effectiveness and robustness of different CNN architectures in solving melanoma diagnosis problem. CNN models can automatically extract high-level features from raw images, overcoming handcrafted features-based methods, and enabling the learning of data-driven features for specific tasks. This type of model can significantly ease decision making process for dermatologists. They can also boost the development of modern tools for a cheaper and better melanoma diagnosis, e.g. by constructing embedded applications in mobile devices that allow an early diagnostic and, therefore, the reduction of invasive treatments and economical resources.

Declaration of Competing Interest

Article title: Convolutional neural networks for the automatic diagnosis of melanoma: an extensive experimental study.

Authors: Eduardo Perez, Oscar Reyes & Sebastian Ventura

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Eduardo Pérez: Formal analysis, Investigation, Software, Validation, Writing - original draft, Writing - review & editing. **Oscar Reyes:** Formal analysis, Investigation, Software, Validation, Writing - original draft, Writing - review & editing. **Sebastián Ventura:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Supervision, Writing - review & editing.

Acknowledgments

This research was supported by the i-PFIS contract no. IFI17/00015 granted by Health Institute Carlos III of Spain. It was also supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund, projects TIN2017-83445-P.

References

Abadi, M., et al., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abbasi, N.R., Shaw, H.M., Rigel, D.S., Friedman, R.J., McCarthy, W.H., Osman, I., Kopf, A.W., Polsky, D., 2004. Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. J. Am. Med. Assoc. 292 (22), 2771–2776.

Abbes, W., Sellami, D., 2017. High-level features for automatic skin lesions neural network based classification. In: Proceedings of the 2nd International Image Processing, Applications and Systems Conference. Hammamet, Tunisia.

Altamura, D., Menzies, S.W., Argenziano, G., Zalaudek, I., Soyer, H.P., Sera, F., Avramidis, M., DeAmbrosis, K., Fargnoli, M.C., Peris, K., 2010. Dermatoscopy of basal cell carcinoma: morphologic variability of global and local features and accuracy of diagnosis. J. Am. Acad. Dermatol. 62 (1), 67–75.

American Cancer Society, 2020. Cancer Facts and Figures. Consulted on April 14, 2020.

Apley, D.W., 2016. Visualizing the effects of predictor variables in black box supervised learning models 44 arXiv:1612.08468.

Argenziano, G., Puig, S., Zalaudek, I., Sera, F., Corona, R., Alsina, M., Barbato, F., Carrera, C., Ferrara, G., Guilabert, A., Massi, D., Moreno-Romero, J.A., Muñoz-Santos, C., Petrillo, G., Segura, S., Soyer, H.P., Zanchini, R., Malvehy, J., 2006. Dermoscopy improves accuracy of primary care physicians to triage lesions suggestive of skin cancer. Journal of Clinical Oncology 24 (12), 1877–1882.

Argenziano, G., Soyer, H.P., De Giorgi, V., 2004. Interactive atlas of dermoscopy. J. Am. Acad. Dermatol. 50 (5), 807–808.

Ballerini, L., Fisher, R.B., Aldridge, B., Rees, J., 2013. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. Lecture Notes in Computational Vision and Biomechanics 6, 63–86.

Baur, C., Albarqouni, S., Navab, N., 2018. Melanogans: high resolution skin lesion synthesis with GANs arXiv:1804.04338.

Ben-David, A., 2008. About the relationship between roc curves and cohen's kappa. Eng. Appl. Artif. Intell. 21 (6), 874–882.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research 13 (Feb), 281–305.

Berthelot, D., Schumm, T., Metz, L., 2017. Began: Boundary equilibrium generative adversarial networks. 1703.10717.

Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using matthews correlation coefficient metric. PLoS ONE 12 (6), e0177678.

Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., 2019. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur. J. Cancer 111, 148–154.

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. In: Proceedings of the 20th International Conference on Pattern Recognition. IEEE, Istanbul, Turkey. 3121–3124

Caruana, R., 1997. Multitask learning. Mach Learn 28 (1), 41–75.

Chin, C.S., Si, J., Clare, A.S., Ma, M., 2017. Intelligent image recognition system for marine fouling using softmax transfer learning and deep convolutional neural networks. Complexity 2017, 5730419:1–5730419:9.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2017). Honolulu, HI, USA, pp. 1800–1807.

Chollet, F., et al., 2015. Keras. https://keras.io.

Ciresan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J., 2010. Deep, big, simple neural nets for handwritten digit recognition. Neural. Comput. 22 (12), 3207–3220.

Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A., 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC-2018). In: Proceedings of the International Symposium on Biomedical Imaging. Vol. 2018-April. Washington, USA, pp. 168–172.

Denton, E., Chintala, S., Szlam, A., Fergus, R., 2015. Deep generative image models using a laplacian pyramid of adversarial networks. Advances in Neural Information Processing Systems. Vol. 2015-Janua. Montreal, Canada.

Dietterich, T., 2000. Ensemble methods in machine learning, 1857 LNCS.

Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115–118.

Ferlay, J., Colombet, M., Soerjomataram, I., Dyba, T., Randi, G., Bettio, M., Gavin, A., Visser, O., Bray, F., 2018. Cancer incidence and mortality patterns in europe: estimates for 40 countries and 25 major cancers in 2018. Eur. J. Cancer 103, 356–387.

Fisher, A., Rudin, C., Dominici, F., 2018. Model class reliance: variable importance measures for any machine learning model class, from the "rashomon" perspective 49 arXiv:1801.01489.

Friedman, M., 1940. A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics 11 (1), 86–92.

Geller, A.C., Swetter, S.M., Brooks, K., Demierre, M.-F., Yaroch, A.L., 2007. Screening, early detection, and trends for melanoma: current status (2000–2006) and future directions. J. Am. Acad. Dermatol. 57 (4), 555–572.

Gilmore, S., Hofmann-Wellenhof, R., Soyer, H., 2010. A support vector machine for decision support in melanoma recognition. Exp. Dermatol. 19 (9), 830–835.

Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M., Petkov, N., 2015. Med-node: a computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Syst. Appl. 42 (19), 6578–6585.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy, pp. 249–256.

Gogul, I., Kumar, V.S., 2017. Flower species recognition system using convolution neural networks and transfer learning. In: Proocedings of the 4th International Conference on Signal Processing, Communication and Networking (ICSCN-2017). Chennai, India.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24 (1), 44–65.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep learning, 1. MIT press Cambridge.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Advances in Neural Information Processing Systems. Vol. 3. Montreal, Quebec, Canada, pp. 2672–2680.

Greenwell, B.M., Boehmke, B.C., McCarthy, A.J., 2018. A simple and effective model-based variable importance measure 27 arXiv:1805.04755.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51 (5), 93.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2019. A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51 (5), 93.

Gutman, D., Codella, N.C.F., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A., 2016. Skin lesion analysis toward melanoma detection: achallenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC-2016) arXiv:1605.01397.

Haenssle, H., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Ben Hadj Hassen, A., Thomas, L., Enk, A., Uhlmann, L., 2018. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology 29 (8), 1836–1842.

Hagerty, J.R., Stanley, R.J., Almubarak, H.A., Lama, N., Kasmi, R., Guo, P., Drugge, R.J., Rabinovitz, H.S., Oliviero, M., Stoecker, W.V., 2019. Deep learning and handcrafted method fusion: higher diagnostic accuracy for melanoma dermoscopy images. IEEE J Biomed Health Inform 23 (4), 1385–1391.

Han, S.S., Kim, M.S., Lim, W., Park, G.H., Park, I., Chang, S.E., 2018. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J. top J, Invest. Dermatol. 138 (7), 1529–1538.

Harangi, B., Baran, A., Hajdu, A., 2018. Classification of Skin Lesions Using An Ensemble of Deep Neural Networks. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. Vol. 2018-July. Honolulu, HI, USA, pp. 2575–2578.

He, H., Garcia, E.A., 2008. Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering (9) 1263–1284.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, pp. 770–778.

Hinton, G., Srivastava, N., Swersky, K., 2012. Rmsprop: divide the gradient by a running average of its recent magnitude. Neural networks for machine learning, Coursera lecture 6e.

Hosseinzadeh Kassani, S., Hosseinzadeh Kassani, P., 2019. A comparative study of deep learning architectures on melanoma detection. Tissue and Cell 58, 76–83.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: efficient convolutional neural networks for mobile vision applications 9 arXiv:1704.04861.

Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., Sun, Q., 2018. Deep learning for image-based cancer detection and diagnosis a survey. Pattern Recognit. 83, 134–149.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K., 2017. Densely connected convolutional networks. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2017). Honolulu, HI, USA.

Huang, L., Zhao, Y.-G., Yang, T.-J., 2019. Skin lesion segmentation using object scale-oriented fully convolutional neural networks. Signal Image Video Process. 13 (3), 431–438.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning, (ICML-2015). Vol. 1. Lille, France, pp. 448–456.

Jaworek-Korjakowska, J., Kleczek, P., Gorgon, M., 2019. Melanoma Thickness Prediction Based on Convolutional Neural Network With VGG-19 Model Transfer Learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, CA, USA.

Jin, L., Gao, S., Li, Z., Tang, J., 2015. Hand-crafted features or machine learnt features? together they improve RGB-D object recognition. In: Proceedings of the IEEE International Symposium on Multimedia (ISM-2014). Taichung, Taiwan, pp. 311–319.

Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G., 2019. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J. Biomed. Health Inform. 23 (2), 538–546.

Keskar, N.S., Socher, R., 2017. Improving generalization performance by switching from adam to sgd 10 arXiv:1712.07628.

Khan, M.A., Javed, M.Y., Sharif, M., Saba, T., Rehman, A., 2019. Multi-model deep neural network based features extraction and optimal selection approach for skin lesion classification. In: Proceedings of the International Conference on Computer and Information Sciences (ICCIS-2019). Karachi, Pakistan.

Kim, H.S., Yoo, K.-Y., Kim, L.H., 2019. Improved performance of image semantic segmentation using nasnet. Korean Chemical Engineering Research 57 (2), 274–282.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization 15 arXiv:1412.6980.

Koh, P.W., Liang, P., 2017. Understanding black-box predictions via influence functions. In: Proceedings of the 34th International Conference on Machine Learning. Vol. 70. Sydney, Australia, pp. 1885–1894.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. Vol. 2. Lake Tahoe, Nevada, USA, pp. 1097–1105.

Laurens van der Maaten, Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2323.

Lee, H.D., Mendes, A.I., Spolaôr, N., Oliva, J.T., Sabino Parmezan, A.R., Wu, F.C., Fonseca-Pinto, R., 2018. Dermoscopic assisted diagnosis in melanoma: reviewing results, optimizing methodologies and quantifying empirical guidelines. Knowl. Based Syst. 158, 9–24.

Li, L., Zhang, Q., Ding, Y., Jiang, H., Thiers, B., Wang, J., 2014. Automatic diagnosis of melanoma using machine learning methods on a spectroscopic system. BMC Med. Imaging 14 (1).

Li, Y., Shen, L., 2018. Skin lesion analysis towards melanoma detection using deep learning network. Sensors (Switzerland) 18 (2).

Liberman, G., Acevedo, D., Mejail, M., 2018. Classification of melanoma images with fisher vectors and deep learning. In: Proceedings of the Iberoamerican Congress on Pattern Recognition, LNCS. Springer, pp. 732–739.

Lin, S., Wang, K., Yang, K., Cheng, R., 2018. Krnet: A kinetic real-time convolutional neural network for navigational assistance. In: Proceedings of the International Conference on Computers Helping People with Special Needs. Springer, Linz, Austria. 55–62

Ling, C., Sheng, V., 2010. Cost-sensitive learning and the class imbalance problem. Encyclopedia of Machine Learning 24, 8.

Liu, X., Wang, X., Matwin, S., 2018. Proceedings of the Interpretable Deep Convolutional Neural Networks via Meta-learning. In: Proceedings of the International Joint Conference on Neural Networks. Vol. 2018-July. Rio de Janeiro, Brazil.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the Advances in Neural Information Processing Systems. Long Beach, CA, USA, pp. 4765–4774.

Mahbod, A., Schaefer, G., Ellinger, I., Ecker, R., Pitiot, A., Wang, C., 2019. Fusing fine-tuned deep features for skin lesion classification. Computerized Medical Imaging and Graphics 71, 19–29.

Matsunaga, K., Hamada, A., Minagawa, A., Koga, H., 2017. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble 4 arXiv:1703.03108.

Mendonca, T., Ferreira, P., Marques, J., Marcal, A., Rozeira, J., 2013. Ph2 - a dermoscopic image database for research and benchmarking. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Osaka, Japan, pp. 5437–5440.

Menegola, A., Tavares, J., Fornaciali, M., Li, L.T., Avila, S., Valle, E., 2017. RECOD Titans at ISIC Challenge 2017 1703.04819.

Nachbar, F., Stolz, W., Merkle, T., Cognetta, A.B., Vogt, T., Landthaler, M., Bilek, P., Braun-Falco, O., Plewig, G., 1994. The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. J. Am. Acad. Dermatol. 30 (4), 551–559.

Nasr-Esfahani, E., Samavi, S., Karimi, N., Soroushmehr, S., Jafari, M., Ward, K., Najarian, K., 2016. Melanoma detection by analysis of clinical images using convolutional neural network. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. Florida, USA. 1373–1376

Perez, F., Vasconcelos, C., Avila, S., Valle, E., 2018. Data Augmentation for Skin Lesion Analysis. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. Springer, Granada, Spain, pp. 303–311.

Perez, L., Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning arXiv:1712.04621.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks 1511.06434.

Rastgoo, M., Lemaître, G., Morel, O., Massich, J., Garcia, R., Mériaudeau, F., Marzani, F., Sidibé, D., 2016. Classification of melanoma lesions using sparse coded features and random forests. In: Progress in Biomedical Optics and Imaging - Proceedings of SPIE. Vol. 9785. San Diego, California, USA.

Reyes, O., Luque, R., Castaño, J., Ventura, S., 2019. A supervised methodology for analyzing dysregulation in splicing machinery: an application in cancer diagnosis. In: Proceedings of the 32nd IEEE CBMS International Symposium on Computer-Based Medical Systems. IEEE, Cordoba, Spain. 120–125

Reyes, O., Ventura, S., 2019. Performing multi-target regression via a parameter sharing-based deep network. Int. J. Neural Syst. 29 (09), 1950014.

Reyes, P., Ventura, S., 2019. Performing multi-target regression via a parameter sharing-based deep network. Int. J. Neural Syst. 1950014.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should {I} Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA. 1135–1144

Rousseeuw, P., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C., 2016. Deepfruits: a fruit detection system using deep neural networks. Sensors (Switzerland) 16 (8).

Sahu, P., Yu, D., Qin, H., 2018. Apply lightweight deep learning on internet of things for low-cost and easy-to-access skin cancer detection. Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, 10579. International Society for Optics and Photonics, Houston, Texas, USA. 1057912

Schwarz, M., Schulz, H., Behnke, S., 2015. RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. In: Proceedings - IEEE International Conference on Robotics and Automation. Vol. 2015-June. Washington, USA, pp. 1329–1335.

Shaffer, J.P., 1986. Modified sequentially rejective multiple test procedures. J. Am. Stat. Assoc. 81 (395), 826–831.

Shapley, L.S., 1953. A value for *n*-person games. Contributions to the Theory of Games 2 (28), 307–317.

Shin, H.-C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-Aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35 (5), 1285–1298.

Siegel, R.L., Miller, K.D., Jemal, A., 2019. Cancer statistics, 2019. CA Cancer Journal for Clinicians 69, 7–34.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition 14 arXiv:1409.1556.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-2017). San Francisco, California, USA, pp. 4278–4284.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 07-12-June-2015. Boston, USA. 1–9

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, pp. 2818–2826.

Thao, L.T., Quang, N.H., 2017. Automatic skin lesion analysis towards melanoma detection. In: Proceedings of the 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems, IES 2017. Vol. 2017-January. Ha Noi, Vietnam. 106–111

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5.

Twinanda, A., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N., 2017. Endonet: A Deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging 36 (1), 86–97.

Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual explanations without opening the black box: automated decisions and the gpdr. Harv. JL & Tech. 31, 841.

Weiss, K., Khoshgoftaar, T.M., Wang, D., 2016. A survey of transfer learning. J. Big Data 3 (1), 9.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics 1 (6), 80–83.

Wilson, A.C., Roelofs, R., Stern, M., Srebro, N., Recht, B., 2017. The marginal value of adaptive gradient methods in machine learning. Adv Neural Inf Process Syst 2017-December (Nips), 4149–4159.

Yoon, C., Hamarneh, G., Garbi, R., 2019. Generalizable Feature Learning in the Presence of Data Bias and Domain Class Imbalance with Application to Skin Lesion Classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Shenzhen, China, pp. 365–373.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Proceedings of the 13th European Conference on Computer Vision. Springer, Zurich, Switzerland. 818–833

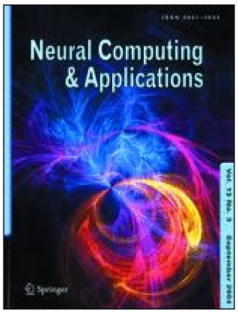
Zeng, G., Zheng, G., 2018. Multi-scale fully convolutional densenets for automated skin lesion segmentation in dermoscopy images. In: Proceedings of the 15th International Conference Image Analysis and Recognition. Springer, Varzim, Portugal. 513–521

Zhao, X.Y., Wu, X., Li, F.-F., Li, Y., Huang, W.H., Huang, K., He, X.Y., Fan, W., Wu, Z., Chen, M.L., Li, J., Luo, Z.L., Su, J., Xie, B., Zhao, S., 2019. The application of deep learning in the risk grading of skin tumors for patients using clinical images. J. Med. Syst. 43 (8).

Zhen, X., Shao, L., Maybank, S.J., Chellappa, R., 2016. Handcrafted vs. learned representations for human action recognition. Image Vis. Comput. 55, 39–41.

Zoph, B., Vasudevan, V., Shlens, J., Le, Q., 2018. Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA. 8697–8710

6.2. An ensemble-based Convolutional Neural Network model powered by a genetic algorithm for melanoma diagnosis



Title	An ensemble-based Convolutional Neural Network model powered by a genetic algorithm for melanoma diagnosis
Authors	E. Pérez, S. Ventura
Journal	Neural Computing and Applications
Volume	Special Issue on Effective and Efficient Deep Learning Based Solutions
Pages	20
Year	2021
Editorial	Springer
DOI	10.1007/s00521-021-06655-7

IF (JCR 2021)	5.102
Category	Computer Science - Artificial Intelligence
Position	45/145 (Q2)

Neural Computing and Applications
https://doi.org/10.1007/s00521-021-06655-7

S. I. : EFFECTIVE AND EFFICIENT DEEP LEARNING

An ensemble-based convolutional neural network model powered by a genetic algorithm for melanoma diagnosis

Eduardo Pérez[1,2] · Sebastián Ventura[1,2,3]

Received: 31 March 2021 / Accepted: 27 October 2021

Abstract

Melanoma is one of the main causes of cancer-related deaths. The development of new computational methods as an important tool for assisting doctors can lead to early diagnosis and effectively reduce mortality. In this work, we propose a convolutional neural network architecture for melanoma diagnosis inspired by ensemble learning and genetic algorithms. The architecture is designed by a genetic algorithm that finds optimal members of the ensemble. Additionally, the abstract features of all models are merged and, as a result, additional prediction capabilities are obtained. The diagnosis is achieved by combining all individual predictions. In this manner, the training process is implicitly regularized, showing better convergence, mitigating the overfitting of the model, and improving the generalization performance. The aim is to find the models that best contribute to the ensemble. The proposed approach also leverages data augmentation, transfer learning, and a segmentation algorithm. The segmentation can be performed without training and with a central processing unit, thus avoiding a significant amount of computational power, while maintaining its competitive performance. To evaluate the proposal, an extensive experimental study was conducted on sixteen skin image datasets, where state-of-the-art models were significantly outperformed. This study corroborated that genetic algorithms can be employed to effectively find suitable architectures for the diagnosis of melanoma, achieving in overall 11% and 13% better prediction performances compared to the closest model in dermoscopic and non-dermoscopic images, respectively. Finally, the proposal was implemented in a web application in order to assist dermatologists and it can be consulted at http://skinensemble.com.

Keywords Convolutional neural networks · Melanoma diagnosis · Ensemble learning · Genetic algorithm · Lesion segmentation

1 Introduction

Melanoma is the most serious form of skin cancer that begins in cells known as melanocytes. Melanoma has an increasing incidence, where just in Europe were estimated 144,200 cases and 20,000 deaths in 2018 [1], whereas in USA, 106,110 new cases of invasive melanoma will be diagnosed (62,260 in men and 43,850 in women) and 7180 deaths are expected in 2021 (4600 men and 2580 women) [2]. The lesion is first diagnosed through an initial clinical screening, and then potentially through a dermoscopic analysis, biopsy and histopathological examination [3]. Despite the expertise of dermatologists, early diagnosis of melanoma remains a challenging task since it is presented in many different shapes, sizes and colors even between samples in the same category [4]. Providing a comprehensive set of tools is necessary for simplifying diagnosis and assisting dermatologists in their decision-making processes [3].

Several automated computer image analysis strategies have been used as tools for medical practitioners to provide accurate lesion diagnostics, including descriptor-based methods [5, 6] and convolutional neural networks (CNNs) [3, 7, 8]. Descriptor-based methods require the previous

✉ Sebastián Ventura
sventura@uco.es

[1] Andalusian Research Institute in Data Science and Computacional Intelligence, DaSCI, University of Córdoba, 14071 Córdoba, Spain

[2] Maimonides Biomedical Research Institute of Cordoba, IMIBIC, University of Córdoba, 14071 Córdoba, Spain

[3] Department of Information Systems, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

Published online: 19 November 2021

extraction of handcrafted features [9], which rely on the expertise of dermatologists and introduce a margin of error. By contrast, CNN models can automatically learn high-level features from raw images [3], thus allowing for the development of applications in a shorter timeframe. Furthermore, Nasr-Esfahani et al. [10] showed that CNN models can overcome handcrafted features-based methods. The authors obtained 7% and 19% better sensitivity performance compared to color-based descriptor and texture-based descriptor, respectively. Recently, Brinker et al. [11] demonstrated that CNN models can match the prediction performance of 145 dermatologists.

CNN models have shown to be effective in solving several complex problems [12, 13]. However, they still present several issues which hamper their accuracy in diagnosing skin conditions. CNN models can learn from a wide variety of nonlinear data points. As such, they are prone to overfitting on datasets with small numbers of samples per category, thus attaining a poor generalization capacity. It is noteworthy that so far, most of the existing public skin datasets only encompass a few hundreds or thousands of images. This can limit the learning capacity of CNN models. On the other hand, CNN models are sensitive to some characteristics in data, such as large inter-class similarities and intra-class variances, variations in viewpoints, changes in lighting conditions, occlusions, and background clutter [14]. These days, the majority of skin datasets are made up of dermoscopic images reviewed by expert dermatologists. However, bear in mind that there is an increased tendency to collect images taken by common digital cameras [15]. The above can reduce invasive treatments and their associated expenses in addition to augmenting the development of modern tools for cheaper and better melanoma diagnoses. Finally, CNN models are approximately invariant with regard to small translations to the input, but they are not rotation, color or lighting-invariant [16, 17]. Invariance is an important concept in the area of image recognition. It means that if you take the input and transform it, the representation you get is the same as the representation of the original. Due to the vast variability in the morphology of moles, this is an important issue to resolve in order to attain a more effective melanoma diagnosis, thus allowing a model to detect rotations or changes to proportion and adapt itself in a way that the learned representation is the same.

Several techniques can be applied to overcome some of these issues, but the most proven include data augmentation [18], transfer learning [3], ensemble learning [19] and, more recently, generative adversarial networks [20] and multi-task learning [21]. However, in most cases, researchers rely on their expertise to select which techniques to apply, and there is no specific pattern to follow that will definitively produce a model with a high level of reliability. Furthermore, most research does not follow a standard experimental study and only includes a limited number of datasets.

As a consequence of the above, in this work, a novel approach for diagnosing melanoma via the use of images is proposed. First, the proposal is inspired by ensemble learning and is built via a genetic algorithm. The genetic algorithm is designed to find the optimal members of such ensembles while considering the entire training phase of the possible members. In addition, the abstract features of all models in the ensemble are merged, resulting in an additional prediction. Next, these individual predictions are combined and a final diagnosis is made. This approach can be seen as a double ensemble, in which all predictive components are double related. The aim is to find the models that best contribute to the ensemble, rather than the individual level. As a result, the state of each CNN model that best trains, generalizes and suits with the other CNN models is selected [22]. In this manner, the training process is implicitly regularized, which has shown better convergence, mitigates the overfitting of the model and improves the generalization performance [19, 23]. To the best of our knowledge, this is the first attempt at intelligently constructing ensembles of CNN models by following a genetic algorithm to better solve the challenge of diagnosing melanoma through the use of images. Second, a novel lesion segmentation method, which is capable of efficiently obtaining reliable segmentation masks without prior information through the use of just a CPU, is applied. Bear in mind that state-of-the-art biomedical segmentation methods commonly require a prior training stage and the use of GPUs. In addition, common techniques such as transfer learning and data augmentation have been applied to further improve performance. To evaluate the suitability of this proposal, an extensive experimental study was conducted on sixteen melanoma-image datasets, enabling a better analysis of the effectiveness of the model. The results showed that the proposed approach achieved promising results and was competitive compared to six state-of-the-art CNN models which have previously been used for diagnosing melanoma [3, 10, 24–26]. These works were summarized in Pérez et al. [8], and the most relevant are explained in the next section. To perform a fair comparison, the above architectures were assessed in Sect. 5.5 by using the same large number of datasets and a quantification of the results was included. Finally, a web application was developed in order to support dermatologists in decision making.

The rest of this work is organized as follows: Sect. 2 briefly presents the state of the art in solving melanoma diagnosis problems mainly by using CNN models; Sect. 3 describes the proposed ensemble convolutional architecture; Sect. 4 describes the genetic algorithm application;

the analysis and discussion of the results are portrayed in Sect. 5; and finally, Sect. 6 outlines conclusions and future works.

2 Related works

The popularity of CNN models increased when AlexNet [27] won the well-known *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) in 2012, reducing the top-5 labels error rate from 26.1% to 15.3%. Since then, CNN models are widely applied in image classification tasks [28–30]. However, most authors apply other sophisticated techniques over CNN models to achieve a better performance in melanoma diagnosis, such as segmentation [31], data augmentation [18], transfer learning techniques [3], and CNN-based ensembles [19].

Data augmentation is employed to add new data to the input space, which helps reducing overfitting [32] and obtaining transformation-invariant models [17]. This technique is usually performed by means of random transformations [33]. Also, most of the datasets available for melanoma diagnosis lack of balance between the categories, so data augmentation can help to tackle the imbalance issue [34]. For example, Esteva et al. [3] showed the suitability of CNN models as a powerful tool for melanoma diagnosis. The authors augmented the images by a factor of 720× using basic transformations, such as rotation, flips and crops. Also, they compared the performance of one CNN model to 21 board-certified dermatologists on biopsy-proven clinical images. The results showed that the CNN model achieved a performance on par with experts. Pérez et al. [8] showed that 12 CNN models achieve better performance when applying data augmentation. For example, Xception [35] increased its average performance by 76%.

Transfer learning has been successfully applied in image classification tasks [3, 30, 36, 37]. For example, Shin et al. [30] applied it specifically in thoraco abdominal lymph node detection and interstitial lung disease classification, in both cases the weights obtained after training with ImageNet were beneficial. Esteva et al. [3] used Google's InceptionV3 [38] architecture pretrained on ImageNet. The authors removed the final classification layer and then they re-trained with 129,450 skin lesions images. Pérez et al. [8] studied the impact of applying transfer learning in skin images, where MobileNet [39] increased its average performance by 42%.

On the other hand, ensemble learning [40] has also shown to be effective in solving complex problems [41, 42]. The combination of several classifiers built from different hypothesis spaces can reach better results than a single classifier [43]. For example, Mahbod et al. [19] proposed an ensemble-based approach where two different CNN architectures were trained with skin lesion images. The results showed to be competitive compared to the state-of-the-art methods for melanoma diagnosis. Harangi et al. [23] proposed an ensemble composed by the well-known CNN architectures AlexNet [27], VGG and InceptionV1 [44]. The ensemble was assessed on the International Symposium on Biomedical Imaging (ISBI) 2017 challenge and obtained very competitive results. However, the ensemble models are usually built trusting in prior knowledge.

In order to achieve better performance, not only new strategies to train the models have been developed, but also efforts have been made to improve the input data. Skin lesion segmentation plays an important role in melanoma diagnosis. It isolates the region of interest and significantly improves the performance of the model. This is a highly complex task, and it is important because some areas not related to the lesion can lead CNN models to misclassify samples. However, some authors decided not to use some of these preprocessing techniques. Mahbod et al. [19] ignored complex preprocessing steps, as well segmentation methods, but applied basic data augmentation techniques to prevent overfitting. Nevertheless, if irrelevant information is removed from the image, the models could be able to achieve better performance. Since 2016, *The International Skin Imaging Collaboration*[1] (ISIC) project annually organizes a challenge in which more than 180 teams have already participated. From 2016 to 2018, there was a special task about lesion segmentation. In ISIC-2016, considering the top performances, the use of segmentation obtained 8% better sensitivity compared to its non-use. Several segmentation methods can be found. For example, Ronneberger et al. [45] designed a CNN model (U-Net) for biomedical image segmentation. The model relies on data augmentation to use the available labeled images more efficiently. The U-Net architecture achieved high performance on different biomedical segmentation applications, such as neuronal structures in electron microscopic recordings and skin lesion segmentation [46, 47]. However, U-Net requires a costly training process using GPU and more important, a prior knowledge from data already segmented by expert dermatologists. Furthermore, Alom et al. [48] proposed a recurrent residual U-Net model (R2U-Net). The model was tested on blood vessel segmentation in retinal images, skin cancer segmentation, and lung lesion segmentation. The results showed better performance compared to the U-Net model. Huang et al. [31] proposed a new segmentation method based on end-to-end object scale-oriented fully convolutional networks. The authors achieved 92.5% of sensitivity and outperformed all

[1] https://www.isic-archive.com.

CNN models in their study, which was 1.4% better compared to the winner in ISIC-2016.

Considering the above, it would be interesting to design a deep learning model that combines features from different approaches such as segmentation, data augmentation, transfer learning and ensemble learning. We hypothesized that evolutionary optimization methods could be an effective approach to find the optimal combination of CNN models [49, 50], considering all states from all the models. Evolutionary methods have proven to be useful in solving many complex problems, finding and optimizing architectures of neural networks [51], and mining imbalanced data [52]. Regarding segmentation methods, in this work it is applied an extension of the Chan-Vese segmentation algorithm [53] and it is evaluated using specialized datasets. To augment data, it is important to perform a data augmentation both on training and test phases [18], which can increase the performance significantly.

3 Ensemble-based convolutional architecture

First of all, the source datasets were preprocessed following an extension of the Chan-Vese segmentation algorithm, as shown in Fig. 1. This algorithm is designed to segment objects without clearly defined boundaries and is based on techniques of curve evolution, Mumford-Shah [54] functional for segmentation and level sets. Chan-Vese has been previously applied in skin image segmentation [55], being an effective starting point to accurately segment skin images. First, Chan-Vese is applied on each input image and as a result, a mask with positive and negative values is obtained. After that, the positive pixels within 40% of the center are selected (cluster of pixels P). After using several recognized skin lesion detection applications such as SkinVision[2], we realized that most of them demand that images must be centered in the lesion in order to perform an accurate diagnosis[3] [56]. In addition, after reviewing a large number of skin image datasets, it was noticed that most images are centered in the lesion, which is an advantage when applying segmentation. Third, all positive clusters (Q) that intersect P are merged, obtaining a new segmentation mask M, $M = Q_1 \cup Q_2 \cup ... \cup Q_n, P \cap Q_i \neq \emptyset$. Finally, the segmented image is obtained after applying the mask M on the original image. In Sect. 5, the proposed segmentation method is compared to other state-of-the-art biomedical segmentation methods to corroborate its effectiveness. Next, the segmented images are used as input to the architecture described below.

Let us say Φ is a model with m independent CNN models, which learn the representations from the same feature space, and an extra prediction block (stacking of dense layers) that yields an extra prediction. The prediction block is obtained by first concatenating all the representations learned by the individual CNN models. Figure 2 shows a specific example of the proposal. An image i is passed as input to the jth CNN model of Φ, and a chromosome indicates which tuples of CNN models and their weights determine the ensemble's architecture. Each CNN outputs the learned representation (r_i^j) and a partial prediction $\hat{o}_i^j$ for the label of this sample by considering the weights required in the chromosome, e.g., the weights from DenseNet trained until epoch 33 ($g_1 = 33$). Thereafter, the representations r_i^j learned by each CNN model are flattened, then concatenated and then passed to the prediction block of the model. A late fusion approach is used to concatenate all the representations learned by the CNN models (Fig. 2c), which has proven to obtain better performance compared to merely combining the individual predictions [57]. Nevertheless, in Sect. 5.5 the prediction block is assessed versus its non-use. Thereafter, we freeze the weights from the individual CNN models and only the prediction block is trained during 20 epochs, providing an additional prediction ($\hat{o}_i^{m+1}$) for the label of the sample. In this work, the prediction block is composed of Dense(512 ReLUs) → Dense(256 ReLUs) → Dense(128 ReLUs) → Dense(1 unit - *sigmoid*). Then, Φ predicts the label of an image by using a soft-voting procedure based on the individual predictions. Regarding efficiency, each CNN model was trained just once and its weights along every epoch of interest were stored in a hard drive.

Once the prediction for a given training sample is computed, the losses produced by each CNN model and the prediction block of Φ are calculated. The loss produced by the jth CNN model of Φ on the ith training image (denoted as $\mathcal{L}^j(i)$) is computed by means of a binary cross-entropy. The goal is to iteratively minimize the prediction errors on the training samples. *Mini-batch Gradient Descent* (MGD) [58] can be applied to solve this optimization problem, since the first derivative of the loss function is well-defined. This algorithm consists in randomly splitting the training set in small batches in order to calculate the model error. The above method has several advantages, such as the model update frequency is higher than batch gradient descent which allows for a more robust convergence, avoiding local minima; batch-based updates provide a computationally more efficient process than stochastic gradient descent and the split in small batches allows the efficiency of not having all training data in memory.

[2] https://www.skinvision.com/.

[3] https://www.firstderm.com/ai-dermatology/.

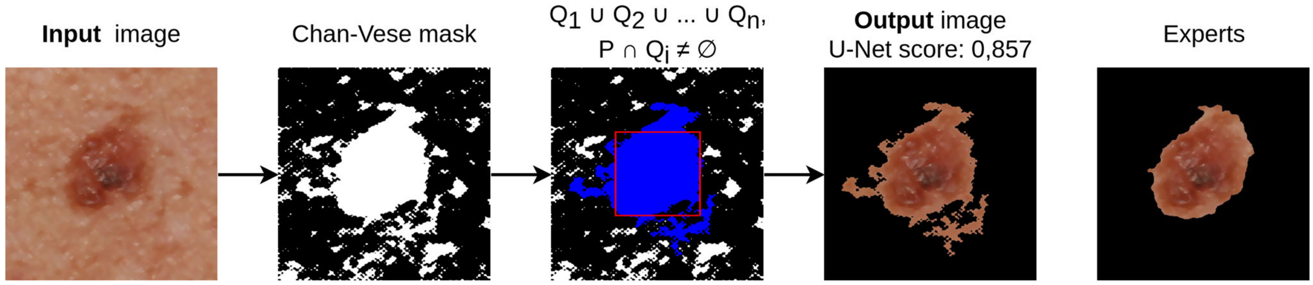


Fig. 1 Preprocessing steps applied before training. U-Net score is the average between Dice coefficient and Jaccard similarity, which have been used before as evaluation metrics in the segmentation task of the ISIC-2018 contest. Input image taken from the DERM-LIB dataset

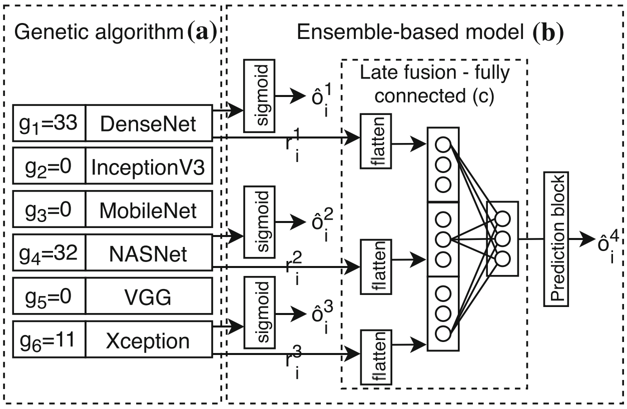


Fig. 2 Example architecture of the proposed ensemble model. Figure shows how a new ensemble is obtained by selecting those CNN models represented by $g_k^j > 0$. In (**a**) the chromosome contains the CNN models and their respective epochs; **b** the ensemble is composed of DenseNet (epoch 33), NASNet (epoch 32) and Xception (epoch 11); **c** a late fusion approach is used to merge all the representations

4 Ensemble model via genetic algorithm

In this work, a genetic algorithm (GA) was designed to find the optimal members of an ensemble model. Next, the different components and steps of the proposed method are explained in detail.

4.1 Individuals and chromosome codification

Let say that the population of the GA has q individuals $\{I_1, I_2, \ldots, I_q\}$, where the j-th individual (I_j) has a chromosome encoded as a list of integer values (from 0 to 150), as shown in Fig. 2a. A chromosome is composed of K genes, where g_k^j represents the k-th gene of the individual I_j. Each gene index k is independent of the others and is related to a specific CNN model. Its value means the epoch of the CNN model to be considered as part of the ensemble; if $g_k^j > 0$, the CNN model is selected; otherwise, it is not selected. Consequently, using this encoding, each individual of the population represents a full solution to the problem.

4.2 Fitness function

The fitness function used to evaluate the individual I_j can be calculated as

$$f_j = \frac{1}{n \times p} \sum_{i=1}^{n} \left[\mathcal{L}_{m+1}^{j}(i) + \sum_{k=1}^{m} \mathcal{L}_k^j(i) \right] \; if \; g_k^j > 0, \qquad (1)$$

where $\mathcal{L}_k^j(i)$ calculates the loss values for the ith image in the CNN model with index k, and $\mathcal{L}_{m+1}^j(i)$ means the loss value of the prediction block; p means the number of times that $g_k^j > 0$ is fulfilled; m represents the total number of CNN models encoded in the chromosome; and n is the total number of images. In summary, for each individual, the fitness function calculates the average loss value between the chosen CNN models and the prediction block; lower average means a more desirable individual.

4.3 Creation of the initial population

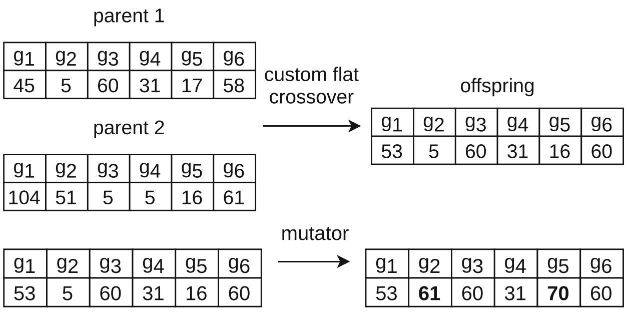
Maintaining a diverse population, especially in the early iterations of the algorithm, is crucial to ensure a good exploration of the search space. The training epochs (e) and the CNN models (m) determine the total number of possible combinations in the search space. To guarantee the diversity, the chromosome of each individual I_j of the population is randomly created, but repeated individuals are not allowed. In this manner, it is possible to avoid the early convergence of the method to local minima. In this work, we have considered an individual as repeated when all genes g_k^j in the chromosome are identical. For example, let us say g_k^A and g_k^B are genes from chromosomes A and B, respectively; k means the index related to a CNN model (e.g., DenseNet, InceptionV3 and Xception) and m means the maximum number of CNN models in each ensemble. A and B are identical if for every $k = 1, 2, ..., m$, $g_1^A = g_1^B$, $g_2^A = g_2^B$, $g_k^A = g_k^B$,..., $g_m^A = g_m^B$. Having said the above, the individuals are repeated if they use a different CNN in any of their genes, i.e., two models are considered different if they have different architectures or the same architecture but with different weights.

4.4 Parent selection

The parents are selected by a tournament selection procedure to create the intermediate population [59]. A tournament size equal 2 was used in this work; the smaller the tournament size, the lower the selection pressure and the search space can be widely considered. To achieve this, two individuals are randomly selected. Then, the individuals are compared and the best individual is selected with replacement, i.e., this individual could be selected in further rounds. This process is repeated until the number of individuals is completed. As the generations increase, the algorithm can focus more on promising regions of the search space.

4.5 Genetic operators

Figure 3 shows the genetic operators applied. First, a custom Flat crossover [60, 61] was performed with a crossover rate p_c, where its general mechanism is adapted for an integer representation scheme. Flat crossover is applied in each pair of genes that are located in the same locus and represents the same CNN model. As a result, a random integer value is chosen from the interval $[g_k^1, g_k^2]$. Once the new offspring is generated, an one-point mutator operator is applied with a probability p_m. The mutation operator randomly selects a gene, then, the value of epoch is changed by a new value selected from a range of possible epochs, e.g., [0; 150]. Finally, the offspring was generated. It is noteworthy that valid individuals are always obtained after performing these operators.



g1	g2	g3	g4	g5	g6
45	5	60	31	17	58

g1	g2	g3	g4	g5	g6
104	51	5	5	16	61

g1	g2	g3	g4	g5	g6
53	5	60	31	16	60

g1	g2	g3	g4	g5	g6
53	5	60	31	16	60

g1	g2	g3	g4	g5	g6
53	**61**	60	31	**70**	60

Fig. 3 Example of the genetic operators applied in the genetic algorithm

4.6 Population update

In this work, a generational elitist algorithm [62] was used to update the population passed from one generation to the next one. As a result, the best individual in the last generation is the best individual of the evolution. To achieve this, the population in each generation keeps all new individuals, as long as the best parent is not better than all the children. In such cases, the best parent replaces the worst child. The worst child is determined by sorting the individuals of the new population according to the fitness value. After replacing the worst child, the new population replaces the previous one. At the end of the algorithm, the best individual in the last generation will be the best ensemble. In this work, the number of individuals generated in each generation is 100 and the population size is kept constant in order to alleviate the computational cost of CNN. However, these values could be tuned depending on the context, including the possibility of increasing/decreasing them. Regarding efficiency, a map with each explored individual and its corresponding fitness value is cached.

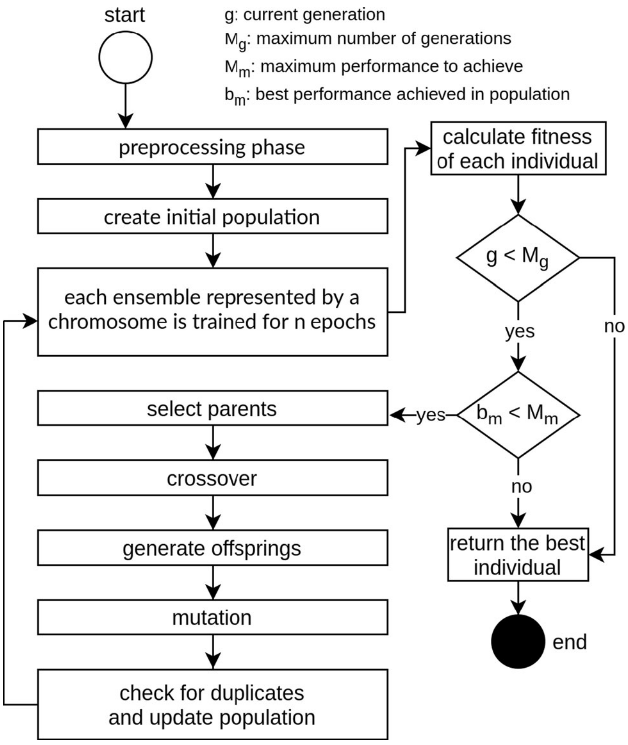

start
g: current generation
M_g: maximum number of generations
M_m: maximum performance to achieve
b_m: best performance achieved in population
preprocessing phase
calculate fitness of each individual
create initial population
$g < M_g$
each ensemble represented by a chromosome is trained for n epochs
yes
no
select parents
yes
$b_m < M_m$
crossover
no
generate offsprings
return the best individual
mutation
end
check for duplicates and update population

Fig. 4 Integration of the genetic algorithm in the training phase

4.7 Integration of the genetic algorithm in the training phase

Figure 4 shows how the training phase was conducted. First, the source images are preprocessed by using the method explained above. Second, given a set of CNN models the initial population of the GA is created. For example, if we have an individual with $g_1^j = 50$ and another one is $g_1^{j+z} = 51$ $(z > 0)$, the last one is obtained after training g_1^j for one more epoch. In general, each CNN model was trained just one time when needed and its weights along every demanded epoch were stored in a hard drive. In the worst case, we train each CNN model 150 epochs once. Third, each ensemble represented by a chromosome is trained for n epochs, bearing in mind that only the prediction block is updated. Then, the fitness of each individual of the current population is calculated by considering the loss obtained in the training images by the members of the ensemble. Four, the parents are selected, then, a crossover and a mutation are performed, and after that a new population is created. Five, the population of the GA is updated. Steps from three to five are repeated until one of the stopping criteria is satisfied. In this work, we have applied the most frequently used stopping criterion which is a specified maximum number of generations. In addition, we stopped the search when an individual had achieved the top performance. Finally, the ensemble model Φ is obtained.

5 Experimental study

This section describes the experimental study carried out in this work. First, the datasets and the experimental protocol are portrayed, and then, the experimental results and a discussion of them are presented.

5.1 Datasets

Table 1 shows a summary of the benchmark datasets. UDA [66], MSK [67], HAM10000 [65] and BCN20000 [63] datasets are included in the ISIC repository. The images are composed strictly of melanocytic lesions that are biopsy-proven and annotated as malignant or benign. Also, images from BCN20000 would be considered hard-to-diagnose and had to be excised and histopathologically diagnosed. PH2[4] [69] comprises dermoscopic images, clinical diagnosis and the identification of several dermoscopic structures. The Dermofit Image Library[5] [64] gathers 1,300 focal high-quality skin lesion images under standardised conditions. Each image has a diagnosis based on expert opinion and like PH2, it includes a binary segmentation mask that denotes the lesion area. MED-NODE[6] [68] collects 170 non-dermoscopic images from common digital cameras; this type of image is very important to prove the models with data from affordable devices. SDC-198[7] [70] contains 6,584 real-world images from 198 categories to encourage further research and its application in real-life scenarios. Finally, DERM7PT[8] [21] is a benchmark dataset composed of clinical and dermoscopic images, allowing to assess how different it is to use dermoscopic images versus images taken with digital cameras.

Only the images labeled as melanoma and nevus were considered, being in total 36,703 images. Most datasets present a high imbalance ratio (ImbR), up to ten times in the case of MSK-3, commonly hampering the learning process. The intra-class (IntraC) and inter-class (InterC) metrics show the average distances between images belonging to different classes, as well as between images belonging to the same class. Both metrics were computed using the Euclidean distance; each image i was represented as a vector. Then, the ratio (DistR) between these metrics showed that both distances are similar, which commonly indicates a high degree of overlapping between classes. Finally, the silhouette score (Silho) [71] was calculated, representing how similar an image is to its own cluster compared to other clusters. The results indicated that images were not well matched to their own cluster, and even samples belonging to different clusters are close in the feature space.

5.2 Experimental settings

Firstly, the proposed segmentation algorithm was applied over PH2 and DERM-LIB, where both datasets were manually segmented by expert dermatologists and they are commonly used as benchmarks due its quality [64, 69]. The aim was to find how much the method is close to expert segmentation. Also, we compared the segmentation algorithm to U-Net and R2U-Net, which are CNN architectures designed for biomedical image segmentation. Bear in mind that the above CNN architectures need prior training. In order to obtain the segmentation mask of an image X, the CNN architectures were trained during 150 epochs with 10% of images as validation set and the rest as training set. The best model obtained during validation was applied on the image X and a segmentation mask was obtained. The

[4] https://www.fc.up.pt/addi/ph2%20database.html.

[5] https://bit.ly/3gi6rKR.

[6] http://www.cs.rug.nl/imaging/databases/melanoma_naevi/.

[7] https://bit.ly/2TdZWQ6.

[8] http://derm.cs.sfu.ca.

Table 1 Summary of the benchmark datasets

Dataset	Source	Img	ImbR	IntraC	InterC	DistR	Silho
BCN20000	[63]	17,393	2.848	9014	10,107	0.892	0.153
DERM-LIB	[64]	407	4.355	7171	9163	0.783	0.270
DERM7PT-C	[21]	827	2.282	15,442	16,318	0.946	0.086
DERM7PT-D	[21]	827	2.282	15,971	16,866	0.947	0.087
HAM10000	[65]	7818	6.024	8705	9770	0.891	0.213
ISBI2016	[66]	1273	4.092	10,553	10,992	0.960	0.101
ISBI2017	[67]	2745	4.259	9280	9674	0.959	0.089
MED-NODE	[68]	170	1.429	9029	9513	0.949	0.068
MSK-1	[67]	1088	2.615	11,753	14,068	0.835	0.173
MSK-2	[67]	1522	3.299	9288	9418	0.986	0.062
MSK-3	[67]	225	10.842	8075	8074	1.000	0.112
MSK-4	[67]	943	3.366	6930	7162	0.968	0.065
PH2	[69]	200	4.000	12,688	14,928	0.850	0.210
SDC-198	[70]	648	4.735	14,054	14,840	0.947	0.116
UDA-1	[66]	557	2.503	11,730	12,243	0.958	0.083
UDA-2	[66]	60	1.609	11,297	11,601	0.974	0.020

above procedure was applied in both PH2 and DERM-LIB datasets.

The second phase aimed at analyzing the GA's hyperparameters on the model's performance. The model was dubbed as *Genetic Algorithm Programming-based Ensemble CNN model* (GAPE-CNN). It should be noted that the main aim of this work was to perform a preliminary study to assess the effectiveness that can be attained by using the proposed architecture in melanoma diagnosis. Consequently, only the hyperparameter values listed in Table 2 values were considered, because including more settings requires high computational resources during training CNN models. Three different number of generations were evaluated, where larger values can lead to a larger number of possible ensembles, but increasing the training cost. Three crossover probabilities were tested, where higher values increase the probability that the genetic information of parents can be combined to generate new individuals. Three mutation probabilities were also evaluated, where higher values allow to escape from local minima and add more exploration of the search space.

In the third phase, the proposal was compared to the following state-of-the-art CNN models that have previously been used in melanoma diagnosis: InceptionV3 [3], DenseNet [25], VGG16 [10], MobileNet [24], Xception [26] and NASNetMobile [72, 73]. Table 2 shows the configuration used to train all the models: the learning rate (α) was equal to 0.01 and it was reduced by a factor of 0.2 if an improvement in predictive performance was not observed during 10 epochs; the weights of the networks were initialized using Xavier method [74] in those cases where transfer learning was not present, e.g., the prediction block and baseline CNN models; a batch of size 8 was used due the medium size of the used datasets and the models were trained along 150 epochs. Mini-batch gradient descent was used for training the models, which is one of the most used optimizers for training CNNs. Despite its simplicity, it performs well across a variety of applications [75] and has been successfully applied for training networks in melanoma diagnosis [7, 18, 76]. In this work, a tuning process was not carried out and so the results could not be conferred to an over-adjustment. The datasets utilized in this work correspond to binary classification problems, so the cost function used for training the models was defined as the average of the binary cross-entropy along all training samples.

Data augmentation technique was mainly applied to tackle the imbalance problem in melanoma diagnosis by applying and combining random rotation-based, flip-based and crop-based transformations over the original images. Bear in mind that color-based transformations were not considered in order not to alter the color space, which is important for the diagnosis of melanoma [77]. As a result, the changes performed during data augmentation do not change the labels of the samples. The data augmentation process was assessed in both training and test data, which can increase the performance significantly [18]. After splitting a dataset into training and test sets, training data

Table 2 Basic configuration used

	Parameter	Value
Segmentation	Threshold	40%
	# epochs (U-Net)	150
Data augmentation	Rotations	[1°,270°]
	Flip	vert. and hori.
	Translations in X, Y	[−30%, 30%]
	Crop	[10%, 30%]
Learning algorithm	# epochs (e)	150
	Mini-batch size	8
	Learning rate (α)	MGD = 0.01
	Momentum (γ)	0.9
Genetic algorithm	Ensemble size (m)	6
	Population size (p)	100
	Top performance (M_m)	1.0
	# generations (M_g)	{100, 250, 400}
	Crossover rate (p_c)	{70%, 80%, 90%}
	Mutation rate (p_m)	{10%, 20%, 30%}

were balanced by creating new images until the number of melanoma images was equal to the normal ones, and the generated training images were considered as independent from the original ones. On the other hand, test data were expanded by randomly augmenting each test image at least ten times, but the generated images remained related to the original ones. Consequently, given an original test image X, the classes' probabilities for X and its related set of images S_X were averaged to yield the final prediction; so any CNN model performs like an ensemble one, where the final probabilities for a test image was computed using a soft-voting strategy.

5.3 Evaluation process

In order to evaluate quantitatively the segmentation method, several performance metrics are considered, including accuracy (ACC), F1-score, Dice coefficient (DC), Jaccard similarity (JS), and the U-Net score (UNS), which is the average between DC and JS. Accuracy and F1-score are calculated using Eqs. 2, 3 and Dice coefficient and Jaccard similarity are calculated using Eqs. 4 and 5.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}, \tag{3}$$

$$DC = 2\frac{|GT \cap SR|}{|GT| + |SR|}, \tag{4}$$

$$JS = \frac{|GT \cap SR|}{|GT \cup SR|}, \tag{5}$$

$$UN = \frac{DC + JS}{2}, \tag{6}$$

where TP, FP, TN and FN represent well-selected pixels, mis-selected pixels, well-discarded pixels and mis-discarded pixels, respectively; GT and SR mean the ground truth pixels and the segmentation result, respectively. The comparison between the segmentation methods is performed using UNS, summarizing DC and JS. The above metrics have been used before in the segmentation task of the ISIC-2018 contest.

Regarding the evaluation process of the CNN models, a 3-times 10-fold cross validation process was performed on the datasets, and the results were averaged across all fold executions. In each fold, *Matthews Correlation Coefficient* (MCC) was used to measure the predictive performance of the models. MCC is widely used in Bioinformatics as a performance metric [78], and it is specially designed to analyze the predictive performance on unbalanced data. MCC is computed as:

$$MCC = \frac{t_p \times t_n - f_p \times f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}}, \tag{7}$$

where t_p, t_n, f_p, and f_n are the number of true positives, true negative, false positives, and false negatives, respectively. MCC value is always in the range $[-1, 1]$, where 1 represents a perfect prediction, 0 indicates a performance similar to a random prediction, and -1 an inverse prediction.

Finally, non-parametric statistical tests were used to detect whether there was any significant difference in predictive performance. Friedman's test [79] was conducted in cases where a multiple comparison was carried out, Hommel's post hoc test [80] was employed to perform a multiple comparison with a control method, Shaffer post hoc test [81] was employed to perform pairwise comparisons, and finally Wilcoxon Signed-Rank test [82] was performed in those cases where only two individual methods were compared. All hypothesis testing was conducted at 95% confidence.

5.4 Software and hardware

The experimental study (training and test) was executed with Ubuntu 18.04, four GPUs NVIDIA Geforce RTX 2080-Ti with 11 GB DDR6 each one and four GPUs NVIDIA Geforce RTX 1080-Ti with 11 GB DDR5X each one. All the experiments were implemented in Python v3.6, and the CNN models were developed by using Keras

Table 3 Summary of the computational resources in deployment time; "D" and "L" mean using default tensorflow-cpu and tensorflow-lite; HDD includes both architecture and weights

CNN	HDD (MB)		RAM (MB)		Time (ms)	
	D	L	D	L	D	L
DenseNet201	141	69	1434	99	255	154
InceptionV3	167	83	974	109	92	111
MobileNet	**25**	**12**	**337**	**40**	**70**	**47**
NASNetMobile	36	16	1126	**40**	156	85
Xception	160	79	778	118	146	200
VGG16	538	537	1229	104	161	398
Total	1,067	796	5878	510	880	995

The best value for each column is highlighted in bold typeface, e.g., MobileNet is the one that consumes less RAM, less hard drive space, and needs less inference time (the aim is to minimize all metrics)

framework v2.2.4 [83] as high level API, and TensorFlow v1.12 [84] as backend. In addition, the proposals were implemented in a web-based application in order to assist dermatologists in decision making. Amazon web services[9] were used in deployment time as platform, providing secure and resizable compute capacity in the cloud. Regarding efficiency, we converted all tensorflow models to Tensorflow Lite models. Tensorflow Lite is a lightweight library for deploying models using a minimum of computational resources. Table 3 shows the minimum amount of hard drive space, RAM and inference time required by each CNN model. The time was measured when processing one sample. Bear in mind that these values were analyzed in deployment time by using CPU and not GPU. As expected, MobileNet was the lightest-weight model which required the least amount of resources. In addition, the highest difference between using or not Tensorflow Lite is regarding RAM—this technology saves more than ten times compared to default deployment. As a result the models only needed 510 MB of RAM and one CPU core to perform inference. In this work was selected basic Amazon EC2 t2.micro, with one GB of RAM and one CPU core ($5.26/month), saving $36.72/month compared to Amazon EC2 t2.large ($41.98/month), which is the closest architecture capable of supporting more than five GB of RAM.

5.5 Results and discussion

In this section, the main results are presented. Additional material for supporting this work can be found at the available web page[10].

5.5.1 Segmentation methods

Table 4 shows the segmentation performance obtained from a small sample (594 images in total). Both datasets contain high-quality images and manual segmentation performed by expert dermatologists. Each image was evaluated applying UNS and other metrics. The proposed segmentation method obtained the best average performance in both datasets with 81% and 80% UNS in PH2 and DERM-LIB, respectively. In addition, Fig. 5 shows Friedman's test ranking, where the proposed method outperformed both U-Net and R2U-Net. Then, Friedman's test rejected the null hypothesis with a p-value $<$ 2.2E-16 in DERM-LIB dataset. The proposal was ranked first, and afterward, the Shaffer's post hoc test was conducted, where the proposal achieved significantly better performance compared to U-Net and R2U-Net. In addition, U-Net significantly outperformed R2U-Net. On the other hand, in PH2 no significant differences were encountered between the three methods, the Friedman's test did not reject the null hypothesis with a p-value equal to 6.476E-1. (The test was conducted with two degrees of freedom, and the Friedman's statistic was equal to 86.898E-2.) However, the proposal was ranked first and it attained 111% and 74% less variance compared to U-Net and R2U-Net, respectively. In this way, the CNN models can focus on relevant pixels, so easing the learning of better abstract and discriminative features for melanoma diagnosis. Also, the proposal did not require prior training, which is a clear advantage compared to those based on CNN models. In addition, the proposed segmentation method can be used with only a CPU, avoiding a significant amount of computational power for training like those using GPU. Finally, R2U-Net obtained a better performance compared to U-Net in PH2, but the opposite occurred in DERMLIB. A larger number of datasets are needed to validate the differences between both methods regarding skin lesion diagnosis.

In addition, we corroborated that the CNN models trained with segmented data were able to achieve better performance. All models achieved the best performance using segmented data, and on average the best ones were MobileNet, DenseNet, and InceptionV3, in that order. Despite MobileNet was designed with efficiency in mind, it managed to overcome more complex CNN models. Nevertheless, DenseNet and InceptionV3 surpassed MobileNet in BCN20000, which is the largest and one of the more complex datasets. Skin image datasets are commonly small, which is where MobileNet usually achieved its best performance [85, 86]. Although these results could be also caused by the use of default hyperparameters, all CNN models shared the same condition in order to avoid advantages between them. On the other hand, the number of epochs could be another cause. This is traditionally large

[9] https://aws.amazon.com/.

[10] http://www.uco.es/kdis/ensemble-based-cnn-melanoma/..

Table 4 Preprocessing performance obtained in DERM-LIB dataset

Image id	U-Net					R2U-Net					Proposed segmentation method				
	ACC	F1	JS	DC	UN	ACC	F1	JS	DC	UN	ACC	F1	JS	DC	UN
P54	0.913	0.892	0.805	0.892	0.849	0.926	0.909	0.833	0.909	0.871	0.968	0.962	0.927	0.962	0.945
P55	0.881	0.725	0.569	0.725	0.647	0.852	0.686	0.522	0.686	0.604	0.895	0.810	0.680	0.810	0.745
P57	0.909	0.856	0.748	0.856	0.802	0.873	0.788	0.650	0.788	0.719	0.964	0.948	0.901	0.948	0.925
P63	0.906	0.857	0.749	0.857	0.803	0.895	0.836	0.718	0.836	0.777	0.950	0.928	0.865	0.928	0.896
P75a	0.791	0.610	0.439	0.610	0.525	0.781	0.680	0.516	0.680	0.598	0.918	0.879	0.784	0.879	0.831
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Average	0.909	0.821	0.726	0.821	0.773	0.885	0.770	0.659	0.770	0.715	0.915	0.844	0.761	0.844	0.803

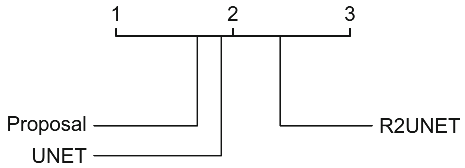


Fig. 5 All pairwise comparisons between the proposed method and the state-of-the-art biomedical segmentation methods in DERM-LIB. The null hypothesis was rejected with adjusted p-value < 2.2E-16

when training CNN models, often hundreds or thousands, allowing the learning algorithm to run until the error from the model has been sufficiently minimized. It is common to find examples in the literature and in tutorials where the number of epochs is set to 500, 1000, and larger. However, in this work we used 150 epochs, mainly because of the number of available images, and even then all models and the proposal achieved competitive performances.

UDA-2 was the most challenging dataset. On average, the models achieved only 44% MCC; UDA-2 has the lowest Silhouette value (0.020), indicating a high overlapping level between classes and increasing the difficulty. The overall best performance was achieved in DERM-LIB, PH2 and HAM10000 datasets with a 90%, 85% and 79% MCC, respectively. The above datasets have the three highest Silhouette values, meaning that they have a low overlapping level between images, which makes easier the task. Table 5 and Fig. 6 summarize the results obtained after training the six CNN models with the segmentation method. The models in Fig. 6 are ordered from left to right according to the ranking computed by Friedman's test, where the proposed segmentation method achieved the best performance in all datasets, followed by TDA. Friedman's test rejected the null hypothesis on all CNN models. It was observed that overall all CNN models using the segmented data presented better results compared to its non-use, proving to be suitable for melanoma diagnosis. Shaffer's post hoc test found that InceptionV3 and Xception significantly improved their performance when segmented data was used compared to the baseline and transfer-learning combined with data augmentation. All baseline CNN models were significantly surpassed by the other techniques. Results showed the proposed segmentation method was able to improve the performance in all CNN models considered in this work. In the following sections, all CNN models are compared using segmented data, which already proved to achieve the best performance.

5.5.2 Analyzing the impact of three main hyperparameters

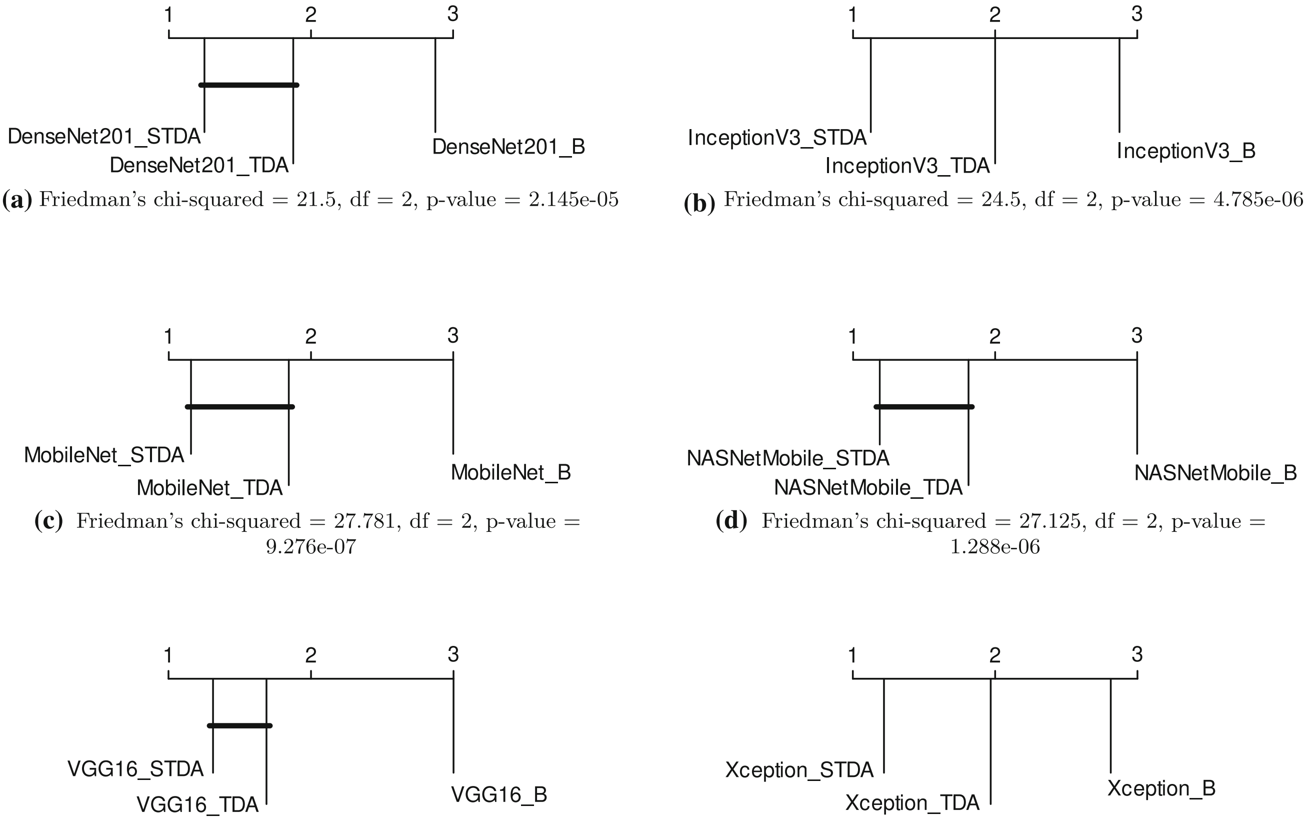
First of all, the advantages of using the extra prediction block were analyzed. The best epochs from each CNN model were combined and all possible ensembles were evaluated. In the end, 57 ensembles were obtained after discarding the empty set and the one-element sets. Then, the best one was selected as baseline (BL) and included in the comparison. In addition, it was obtained another baseline model applying the same above procedure by simply combining the predictions of the CNN models without using the prediction block (BLs). Table 6 shows the average MCC values on test sets comparing BL versus BLs. Results show that models using the extra prediction block obtained the best performance in all datasets compared to BLs. Finally, the Wilcoxon's test rejected the null hypothesis with a p-value equal to 2.189E-4, confirming the benefit and effectiveness of using the prediction block for combining the features from the individual CNN models. Henceforth, the rest of the experimental study was executed using the proposal architecture.

Table 7 shows the average MCC results attained by the proposed genetic algorithm using different values of g, p_c and p_m. For each dataset, the best MCC value is highlighted in bold typeface. Overall, all models obtained a high performance, being the lowest and the highest average performance 96.9% and 97.8%, respectively. The models attained their best performance in MSK-3, PH2, DERM-LIB and SDC-198 datasets. Regarding the parameters that

Table 5 Average MCC values obtained by using six state-of-the-art CNN models.

Dataset	DenseNet			InceptionV3			MobileNet			NASNet			VGG16			Xception		
	B	TDA	STDA	B	TDA	STDA	B	TDA	STDA	B	TDA	STDA	B	TDA	STDA	B	TDA	STDA
BCN20000	0.642	0.929	0.933	0.633	0.934	**0.938**	0.588	0.919	0.928	0.661	0.915	0.916	0.494	0.705	0.809	0.603	0.918	0.919
DERM-LIB	0.888	0.934	0.993	0.807	0.936	0.965	0.859	0.954	0.992	0.853	0.885	0.978	0.719	0.724	0.941	0.934	0.918	0.977
DERM7PT-C	0.389	0.803	0.813	0.317	0.728	0.771	0.315	0.806	0.812	0.296	0.723	0.728	0.000	**0.877**	0.722	0.299	0.740	0.741
DERM7PT-D	0.502	0.890	0.848	0.455	0.838	0.843	0.458	0.893	**0.903**	0.406	0.853	0.854	0.321	0.851	0.738	0.488	0.849	0.849
HAM10000	0.628	0.954	**0.960**	0.591	0.940	0.942	0.463	0.945	0.947	0.578	0.934	0.935	0.383	0.763	0.835	0.479	0.959	0.894
ISBI2016	0.367	0.850	**0.878**	0.307	0.802	0.825	0.309	0.850	0.854	0.254	0.805	0.829	0.000	0.557	0.667	0.281	0.799	0.843
ISBI2017	0.217	0.854	0.864	0.225	0.829	0.839	0.095	0.875	0.875	0.230	0.832	0.849	0.176	0.743	0.805	0.193	0.846	**0.882**
MED-NODE	0.508	0.698	0.699	0.567	0.732	**0.766**	0.533	0.741	0.768	0.466	0.660	0.665	0.000	0.611	0.675	0.539	0.745	0.759
MSK-1	0.555	0.880	**0.940**	0.574	0.868	0.873	0.494	0.886	0.899	0.533	0.843	0.852	0.487	0.667	0.683	0.405	0.856	0.860
MSK-2	0.325	0.830	**0.889**	0.282	0.805	0.817	0.263	0.860	0.870	0.269	0.785	0.867	0.239	0.563	0.581	0.232	0.815	0.832
MSK-3	0.165	1.000	0.959	0.080	0.959	0.969	0.124	**1.000**	0.966	0.119	**1.000**	0.808	0.000	0.907	0.814	0.149	0.927	0.938
MSK-4	0.393	0.864	0.868	0.365	0.844	0.847	0.269	0.890	0.850	0.302	0.857	0.823	0.000	**0.906**	0.799	0.273	0.822	0.829
PH2	0.771	0.960	0.967	0.647	0.963	0.964	0.700	0.963	**0.987**	0.640	0.934	0.937	0.317	0.909	0.911	0.808	0.944	0.971
SDC-198	0.610	0.951	0.956	0.444	0.908	0.929	0.496	0.971	0.979	0.502	0.889	0.922	0.000	**0.980**	0.836	0.456	0.930	0.941
UDA-1	0.394	0.764	0.808	0.336	0.720	0.762	0.360	0.781	**0.813**	0.351	0.706	0.716	0.286	0.601	0.690	0.348	0.692	0.734
UDA-2	**0.596**	0.522	0.513	0.540	0.413	0.410	0.375	0.577	0.580	0.427	0.548	0.441	0.000	0.425	0.437	0.439	0.362	0.355

The best MCC values by dataset were highlighted in bold typeface. “B” means a baseline performance (i.e., none technique was applied). “TDA” and “STDA” represent the model applying transfer learning and data augmentation, and combining segmented images with transfer learning and data augmentation, respectively; data augmentation was applied both in training and test sets.



(a) Friedman's chi-squared = 21.5, df = 2, p-value = 2.145e-05

(b) Friedman's chi-squared = 24.5, df = 2, p-value = 4.785e-06

(c) Friedman's chi-squared = 27.781, df = 2, p-value = 9.276e-07

(d) Friedman's chi-squared = 27.125, df = 2, p-value = 1.288e-06

(e) Friedman's chi-squared = 25.125, df = 2, p-value = 3.501e-06

(f) Friedman's chi-squared = 20.344, df = 2, p-value = 3.823e-05

Fig. 6 Each sub-figure shows the methods ordered from left to right according to the ranking computed by Friedman's test; the proposed segmentation method achieved the best performance in all CNN models. The lines located summarizes the significant differences encountered by the Shaffer's post hoc test, in such a way that groups of models that are not significantly different (at α = 0.05) are connected by a line

control the GA, the results showed the higher the number of generations, the better was the predictive performance. A higher number of generations could imply more exploration and higher possibilities to create offspring with a more accurate set of CNN models. Furthermore, the best results peaked when using a high mutation probability, denoting that a higher mutation probability could imply a better exploration of good sets of epochs. Finally, it was observed that 250 generations are good enough to obtain a performance which is significantly similar to the top performance obtained by using 400 generations.

The Rank column shows the average ranking computed by Friedman's test, and this ranking found that $g = 400$, $p_c = 90\%$ and $p_m = 30\%$ were the best settings for the GA, obtaining on average the best predictive performance. The Friedman's test rejected the null hypothesis with a p-value equal to 8.771E-15; Friedman's statistic was equal to 124.28 with 26 degrees of freedom. Afterward, the Hommel's post hoc test was conducted by considering the GA as the control method, which significantly outperformed 50% of all other considered configurations. Next, the best GA is compared to several state-of-the-art CNN models that have previously been used in melanoma diagnosis and at the same time are the baseline to build the ensemble.

5.5.3 Comparing with state-of-the-art CNN models

Table 8 shows the fold changes between the proposed model and each of the state-of-the-art CNN models. The BL ensemble overcame all individual state-of-the-art models, except in DERM-LIB, where DenseNet201 and MobileNet surpassed it by a small margin. It should be noticed that these models are the same considered by the GA to build the ensemble, so this comparison plays another role, which is to corroborate that ensemble learning is more suitable for melanoma diagnosis tasks compared to individual models. The results were very promising since the proposal GAPE-CNN achieved the highest MCC values in all datasets. It is noteworthy that the proposal achieved a predictive performance 165% and 130% higher than Xception and InceptionV3 on UDA-2 dataset, respectively. Also, it achieved high performance in the two largest

Table 6 Average MCC values on test sets; BL represents the best ensemble obtained by combining the individual best models from each architecture and BLs represents the same as BL, but simply merging the predictions without using the extra prediction block

Dataset	BLs	BL
BCN20000	0.934	**0.949**
DERM-LIB	0.982	**0.990**
DERM7PT-C	0.852	**0.868**
DERM7PT-D	0.916	**0.929**
HAM10000	0.963	**0.970**
ISBI2016	0.911	**0.933**
ISBI2017	0.919	**0.963**
MED-NODE	0.808	**0.890**
MSK-1	0.943	**0.953**
MSK-2	0.902	**0.925**
MSK-3	0.960	**0.970**
MSK-4	0.938	**0.961**
PH2	0.960	**0.980**
SDC-198	0.955	**0.995**
UDA-1	0.822	**0.890**
UDA-2	0.657	**0.776**

The best MCC values are highlighted in bold typeface. Wilcoxon's test rejected the null hypothesis with a p-value equal to 2.189E-4

datasets. The best average predictive performance was observed on DERM-LIB and PH2 datasets, where it was obtained 98% and 96% MCC values, respectively. The lowest overall predictive performance was observed on the dataset UDA-2 with 56% of MCC. However, the proposal was at least 62% and 21% better than all the individual CNN models and the BL, respectively. Overall, the worst performance was attained by VGG16.

Table 9 summarizes the average MCC values on test sets comparing the baseline versus GAPE-CNN. The proposal surpassed BL in all datasets, and finally, the Wilcoxon's test rejected the null hypothesis with a p-value equal to 2.189E-4, confirming the benefit and effectiveness of using genetic algorithms for learning the set of CNN models to build an ensemble. Each independent CNN model provides a partial prediction, which is aggregated to yield a final decision. The architecture follows an ensemble approach, which has demonstrated to be an effective way to improve the learning process in many real-world problems [87]. Furthermore, the proposed model applies transfer learning and data augmentation techniques. Data are augmented not only at training stage, but also at test stage, thus allowing to attain a better predictive performance. The data augmentation process on both phases has shown to be an effective way to improve melanoma diagnosis [88], and also it is an excellent approach to cope with the imbalance data issue, and the high inter- and intra-class variability present in most skin image datasets.

5.5.4 Dermoscopic versus non-dermoscopic images

Figure 7 shows the average performance attained by the models by grouping the datasets in dermoscopic and non-dermoscopic ones. The results showed that the proposed architecture attained the best performance whatever the type of image, denoting the effectiveness of the approach. All CNN models attained their best performance using dermoscopic images by a slight margin. NASNetMobile was the architecture most benefited from using dermoscopic images, with an improvement of about 8%. The biggest improvement was found comparing with VGG; the proposal outperformed VGG considering dermoscopic and non-dermoscopic images in 31% and 30%, respectively. The second best performance behind the proposal was achieved by the BL in both types of images; the proposal surpassed it by 5% in both types of images. To sum up, the results obtained through the experimental study revealed that GAPE-CNN was effective for diagnosing melanoma, attaining better predictive performance with respect to the state-of-the-art models.

6 Conclusions

It is clear that automatic melanoma diagnosis via deep learning models is a challenging task, mainly due to a lack of data and differences even between samples from the same category. We addressed these problems via a series of contributions. First, to preprocess data, we validated and applied an extension of the Chan-Vese algorithm. The segmentation masks indicated that the proposal achieved a better performance compared to state-of-the-art segmentation methods. Also, the results showed that all CNN models improved their performance by using segmented data. Second, the training and testing data were enriched using data augmentation, reducing overfitting and obtaining transformation-invariant models. Third, we increased the performance by applying transfer learning from the pre-trained ImageNet. Also, transfer learning alleviated the requirement for a large number of training data. Finally, to further improve the discriminative power of CNN models, we proposed a novel ensemble method based on a genetic algorithm, which finds an optimal set of CNN models. An extensive experimental study was conducted on sixteen image datasets, demonstrating the utility and effectiveness of the proposed approach, attaining a high predictive performance even in datasets with complex properties. Results also showed the proposed ensemble model is competitive with regard to state-of-the-art computational methods.

Table 7 Average MCC values on test sets; g, p_c and p_m represent generations, crossover rate and mutation rate, respectively

g	p_c	p_m	BCN	DERM	DER-C	DER-D	HAM	IS-16	IS-17	MED	MS-1	MS-2	MS-3	MS-4	PH2	SDC	UD-1	UD-2	Rank	p-value
100	0.7	0.1	0.977	**1.000**	0.926	0.944	0.997	0.990	0.992	0.931	0.977	0.986	**1.000**	0.979	**1.000**	**1.000**	**0.939**	0.915	1.600E+01	1.877E-02
		0.2	0.977	**1.000**	0.915	0.945	0.997	0.991	0.991	0.942	0.979	0.986	**1.000**	0.976	**1.000**	**1.000**	0.935	0.915	1.703E+01	6.271E-03
		0.3	0.978	**1.000**	0.918	0.950	0.998	0.991	0.993	0.931	0.977	0.984	**1.000**	0.976	**1.000**	**1.000**	0.935	0.915	1.575E+01	2.549E-02
	0.8	0.1	0.978	**1.000**	0.920	0.942	0.997	0.989	0.992	0.920	0.977	0.985	**1.000**	0.977	**1.000**	0.995	0.934	0.912	2.000E+01	8.308E-05
		0.2	0.976	0.993	0.917	0.947	0.997	0.990	0.992	0.931	0.979	0.989	**1.000**	0.977	**1.000**	**1.000**	0.935	0.915	1.719E+01	5.295E-03
		0.3	0.976	**1.000**	0.918	0.945	0.998	0.989	0.993	0.931	0.977	0.985	**1.000**	0.976	**1.000**	**1.000**	**0.939**	**0.941**	1.606E+01	1.737E-02
	0.9	0.1	0.977	0.992	0.912	0.945	0.997	0.988	0.993	0.919	0.975	0.984	**1.000**	0.973	**1.000**	0.995	0.931	0.915	2.231E+01	1.262E-06
		0.2	0.975	0.993	0.912	0.944	0.997	0.993	0.991	0.920	0.977	0.984	**1.000**	0.979	**1.000**	0.995	0.931	0.915	2.119E+01	1.072E-05
		0.3	0.976	0.993	0.917	0.939	0.997	0.986	0.991	0.931	0.975	0.984	**1.000**	0.976	**1.000**	**1.000**	0.935	**0.941**	2.056E+01	3.228E-05
250	0.7	0.1	0.979	**1.000**	0.923	0.948	0.999	0.993	0.995	0.942	0.979	0.987	**1.000**	0.982	**1.000**	**1.000**	**0.939**	**0.941**	8.250E+00	~~6.560E-01~~
		0.2	0.979	**1.000**	0.923	0.947	0.998	0.993	0.992	0.942	0.979	0.986	**1.000**	0.982	**1.000**	**1.000**	**0.939**	**0.941**	1.041E+01	~~6.560E-01~~
		0.3	0.978	**1.000**	0.920	0.948	0.998	0.993	0.993	0.931	0.979	0.987	**1.000**	**0.988**	**1.000**	0.995	**0.939**	**0.941**	1.138E+01	~~6.220E-01~~
	0.8	0.1	0.978	0.993	0.923	0.945	0.998	0.993	0.995	0.931	0.979	0.986	**1.000**	0.977	**1.000**	**1.000**	0.931	**0.941**	1.388E+01	~~1.858E-01~~
		0.2	0.978	**1.000**	0.920	0.945	0.998	0.993	**0.997**	0.943	0.979	0.986	**1.000**	0.979	**1.000**	**1.000**	0.935	0.915	1.228E+01	~~5.386E-01~~
		0.3	0.978	**1.000**	0.923	0.948	0.998	0.989	0.993	0.931	0.979	0.989	**1.000**	0.982	**1.000**	**1.000**	**0.939**	**0.941**	1.106E+01	~~6.479E-01~~
	0.9	0.1	0.977	**1.000**	0.920	0.947	0.997	0.991	0.992	0.931	0.979	0.984	**1.000**	0.979	**1.000**	0.995	0.935	**0.941**	1.628E+01	1.365E-02
		0.2	0.977	**1.000**	0.918	0.942	0.998	0.991	0.994	0.931	0.977	0.986	**1.000**	0.979	**1.000**	**1.000**	0.935	0.915	1.600E+01	1.877E-02
		0.3	0.977	0.993	0.917	0.948	0.997	0.988	0.994	0.920	0.977	0.984	**1.000**	0.979	**1.000**	**1.000**	0.935	0.915	1.816E+01	1.545E-03
400	0.7	0.1	0.979	**1.000**	0.926	0.950	0.998	0.993	0.994	0.942	**0.982**	0.987	**1.000**	0.980	**1.000**	**1.000**	0.935	**0.941**	8.812E+00	~~6.560E-01~~
		0.2	0.979	**1.000**	0.929	0.947	0.998	0.993	0.993	0.943	0.979	0.989	**1.000**	0.979	**1.000**	**1.000**	0.935	**0.941**	9.875E+00	~~6.560E-01~~
		0.3	0.979	**1.000**	0.923	0.947	0.999	0.991	0.995	0.931	0.979	**0.991**	**1.000**	**0.988**	**1.000**	**1.000**	**0.939**	**0.941**	9.031E+00	~~6.560E-01~~
	0.8	0.1	0.979	**1.000**	0.922	0.947	0.998	0.993	0.995	**0.954**	0.979	0.988	**1.000**	0.979	**1.000**	**1.000**	**0.939**	**0.941**	9.250E+00	~~6.560E-01~~
		0.2	0.979	**1.000**	0.923	0.945	0.998	0.991	0.996	0.931	0.979	0.989	**1.000**	0.982	**1.000**	0.995	**0.939**	**0.941**	1.103E+01	~~6.479E-01~~
		0.3	0.979	**1.000**	0.929	0.945	0.998	0.993	0.995	0.942	0.979	0.986	**1.000**	0.982	**1.000**	**1.000**	**0.939**	0.915	1.025E+01	~~6.560E-01~~
	0.9	0.1	0.978	**1.000**	0.926	0.947	0.997	0.993	0.993	**0.954**	0.979	0.986	**1.000**	0.977	**1.000**	0.995	0.935	**0.941**	1.300E+01	~~3.590E-01~~
		0.2	0.978	**1.000**	0.920	0.945	0.997	0.993	0.991	0.942	0.979	0.984	**1.000**	0.979	**1.000**	**1.000**	0.931	0.915	1.597E+01	1.951E-02
		0.3	**0.982**	**1.000**	**0.940**	**0.955**	**1.000**	**0.995**	0.993	**0.954**	**0.982**	0.984	**1.000**	**0.988**	**1.000**	**1.000**	**0.939**	**0.941**	**7.000E+00**	–

Rank means the average ranking computed by Friedman's test; Hommel's p-values are shown in the last column where strike values represent the ones do not have significant differences compared to the best ranked (last row). The best MCC values for each dataset are highlighted in bold typeface; dataset names were shortened and placed horizontally to save space

Table 8 Average MCC values of the different models on each dataset

Dataset	DenseNet201	%	InceptionV3	%	MobileNet	%	NASNetMobile	%	VGG16	%	Xception	%	BL	%	GAPE-CNN
BCN20000	0.933	5	0.938	5	0.928	6	0.916	7	0.809	21	0.919	7	0.949	3	**0.982**
DERM-LIB	0.993	1	0.965	4	0.992	1	0.978	2	0.941	6	0.977	2	0.990	1	**1.000**
DERM7PT-C	0.813	16	0.771	22	0.812	16	0.728	29	0.722	30	0.741	27	0.868	8	**0.940**
DERM7PT-D	0.848	13	0.843	13	0.903	6	0.854	12	0.738	29	0.849	12	0.929	3	**0.955**
HAM10000	0.960	4	0.942	6	0.947	6	0.935	7	0.835	20	0.894	12	0.970	3	**1.000**
ISBI2016	0.878	13	0.825	21	0.854	17	0.829	20	0.667	49	0.843	18	0.933	7	**0.995**
ISBI2017	0.864	15	0.839	18	0.875	13	0.849	17	0.805	23	0.882	13	0.963	3	**0.993**
MED-NODE	0.699	36	0.766	25	0.768	24	0.665	43	0.675	41	0.759	26	0.890	7	**0.954**
MSK-1	0.940	4	0.873	12	0.899	9	0.852	15	0.683	44	0.860	14	0.953	3	**0.982**
MSK-2	0.889	11	0.817	20	0.870	13	0.867	13	0.581	69	0.832	18	0.925	6	**0.984**
MSK-3	0.959	4	0.969	3	0.966	4	0.808	24	0.814	23	0.938	7	0.970	3	**1.000**
MSK-4	0.868	14	0.847	17	0.850	16	0.823	20	0.799	24	0.829	19	0.961	3	**0.988**
PH2	0.967	3	0.964	4	0.987	1	0.937	7	0.911	10	0.971	3	0.980	2	**1.000**
SDC-198	0.956	5	0.929	8	0.979	2	0.922	8	0.836	20	0.941	6	0.995	1	**1.000**
UDA-1	0.808	16	0.762	23	0.813	15	0.716	31	0.690	36	0.734	28	0.890	6	**0.939**
UDA-2	0.513	83	0.410	130	0.580	62	0.441	113	0.437	115	0.355	165	0.776	21	**0.941**

The % columns represent the fold changes after comparing the proposed model with the others, e.g., GAPE-CNN attained a 21% of improvement over VGG16 on BCN20000 dataset. The best MCC values for each dataset are highlighted in bold typeface

Table 9 Average MCC values on test sets; BL represents the best ensemble obtained by combining the individual best model from each architecture and GAPE-CNN represents our proposal

Dataset	BL	GAPE-CNN
BCN20000	0.949	**0.982**
DERM-LIB	0.990	**1.000**
DERM7PT-C	0.868	**0.940**
DERM7PT-D	0.929	**0.955**
HAM10000	0.970	**1.000**
ISBI2016	0.933	**0.995**
ISBI2017	0.963	**0.993**
MED-NODE	0.890	**0.954**
MSK-1	0.953	**0.982**
MSK-2	0.925	**0.984**
MSK-3	0.970	**1.000**
MSK-4	0.961	**0.988**
PH2	0.980	**1.000**
SDC-198	0.995	**1.000**
UDA-1	0.890	**0.939**
UDA-2	0.776	**0.941**

The best MCC values are highlighted in bold typeface. Wilcoxon's test rejected the null hypothesis with a p-value equal to 2.189E-4

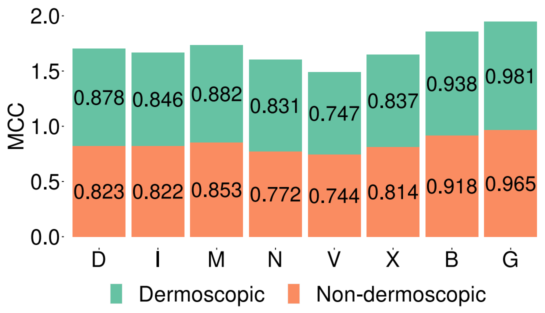


Fig. 7 Average MCC values on test sets by grouping the datasets in dermoscopic (green bars) and non-dermoscopic (orange bars); "D": DenseNet201; "I": InceptionV3; "M": MobileNet; "N": NASNetMobile; "V": VGG16; "X": Xception; "B": base line; "G": our proposal

Future works will conduct more extensive experiments to validate the full potential of the proposed architecture, for example by considering a wide set of hyperparameters to be tuned as well as a larger number of datasets. Finally, it is noteworthy that our approach benefits from combining different CNN models—abstract features are combined in the extra prediction block, and individual predictions from the ensemble and the mentioned block are combined to obtain a diagnosis. This approach is not restricted to melanoma diagnosis problems and could be applied on other real-world problems in the future.

Ackowledgements This research was supported by the i-PFIS contract no. IFI17/00015 granted by Health Institute Carlos III of Spain. It was also supported by the University of Cordoba and the European Regional Development Fund, project UCO-FEDER 18 REF.1263116 MOD.A.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Consejería de Transformación Económica, Industria, Conocimiento y Universidades. Junta de Andalucía (ES) (Grant No. UCO-1262678) and ministerio de ciencia e innovación (Grant No. PID2020-115832GB-I00).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, Gavin A, Visser O, Bray F (2018) Cancer incidence and mortality patterns in Europe: estimates for 40 countries and 25 major cancers in 2018. Eur J Cancer 103:356–387
2. American Cancer Society: Cancer Facts and Figures (2021). https://bit.ly/3gNDBVr. Consulted on June 22, 2021
3. Esteva A, Kuprel B, Novoa R, Ko J, Swetter S, Blau H, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118
4. Geller AC, Swetter SM, Brooks K, Demierre MF, Yaroch AL (2007) Screening, early detection, and trends for melanoma: current status (2000–2006) and future directions. J Am Acad Dermatol 57(4):555–572
5. Rastgoo M, Lemaître G, Morel O, Massich J, Garcia R, Mériaudeau F, Marzani F, Sidibé D (2016) Classification of melanoma lesions using sparse coded features and random forests. In: progress in biomedical optics and imaging - proceedings of SPIE, vol. 9785. San Diego, California, USA
6. Sánchez-Monedero J, Pérez-Ortiz M, Sáez A, Gutiérrez PA, Hervás-Martínez C (2018) Partial order label decomposition approaches for melanoma diagnosis. Appl Soft Comput 64:341–355
7. Li X, Yu L, Fu C.W., Heng P.A. (2018) Deeply supervised rotation equivariant network for lesion segmentation in dermoscopy images. Lect Notes Comput Sc 11041 LNCS, 235–243
8. Pérez E, Reyes O, Ventura S (2021) Convolutional neural networks for the automatic diagnosis of melanoma: an extensive experimental study. Med Image Anal 67:101858
9. Jin L, Gao S, Li Z, Tang J (2015) Hand-crafted features or machine learnt features? together they improve RGB-D object

recognition. In: proceedings of the IEEE ISM-2014, pp. 311–319. Taichung, Taiwan

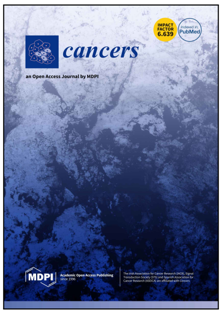
10. Nasr-Esfahani E, Samavi S, Karimi N, Soroushmehr S, Jafari M, Ward K, Najarian K (2016)Melanoma detection by analysis of clinical images using convolutional neural network. In: proceedings of the IEEE EMBS, pp. 1373–1376. Florida, USA
11. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A (2019) A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Cancer 111:148–154
12. Liu Y, Chen X, Peng H, Wang Z (2017) Multi-focus image fusion with a deep convolutional neural network. Inform Fusion 36:191–207
13. Hossain MS, Muhammad G (2019) Emotion recognition using deep learning approach from audio-visual emotional big data. Inform Fusion 49:69–78
14. Asif U, Bennamoun M, Sohel F (2018) A multi-modal, discriminative and spatially invariant CNN for RGB-D object labeling. IEEE T Pattern Anal 40(9):2051–2065
15. Ericsson: on the pulse of the networked society. Tech. rep. (2015). https://apo.org.au/node/59109
16. Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. arXiv:1710.09829
17. Lenc K, Vedaldi A (2019) Understanding image representations by measuring their equivariance and equivalence. Int J Comput Vision 127(5):456–476
18. Perez F, Vasconcelos C, Avila S, Valle E (2018) Data augmentation for skin lesion analysis. In: OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis, pp. 303–311. Springer, Granada, Spain
19. Mahbod A, Schaefer G, Ellinger I, Ecker R, Pitiot A, Wang C (2019) Fusing fine-tuned deep features for skin lesion classification. Comput Med Imag Grap 71:19–29
20. Baur C, Albarqouni S, Navab N (2018) MelanoGANs: high resolution skin lesion synthesis with GANs. arXiv preprint: arXiv:1804.04338
21. Kawahara J, Daneshvar S, Argenziano G, Hamarneh G (2019) Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J Biomed Health 23(2):538–546
22. Zeng H, Haleem H, Plantaz X, Cao N, Qu H (2017) CNNComparator: Comparative Analytics of Convolutional Neural Networks. arXiv:1710.05285
23. Harangi B, Baran A, Hajdu A (2018) Classification of skin lesions using an ensemble of deep neural networks. In: proceedings of the annual international conference of the IEEE EMBS, vol. 2018-July, pp. 2575–2578. Honolulu, HI, USA
24. Sahu P, Yu D, Qin H (2018) Apply lightweight deep learning on internet of things for low-cost and easy-To-Access skin cancer detection. In: progress in biomedical optics and imaging - proceedings of SPIE, vol. 10579. Houston, Texas, USA
25. Zeng G, Zheng G (2018) Multi-scale fully convolutional dense-Nets for automated skin lesion segmentation in dermoscopy images. Lect Notes Comput Sci 10882 LNCS, 513–521
26. Zhao XY, Wu X, Li FF, Li Y, Huang WH, Huang K, He XY, Fan W, Wu Z, Chen ML, Li J, Luo ZL, Su J, Xie B, Zhao S (2019) The application of deep learning in the risk grading of skin tumors for patients using clinical images. J Med Syst 43(8):1–7
27. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst, vol 2. Harrahs and Harveys, Lake Tahoe, NV, USA, pp 1097–1105
28. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li F.F. (2014) Large-scale video classification with convolutional neural networks. In: proceedings of the IEEE computer society CVPR, pp. 1725–1732. Washington, USA
29. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE T Pattern Anal 37(9):1904–1916
30. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE T Med Imaging 35(5):1285–1298
31. Huang L, Zhao YG, Yang TJ (2019) Skin lesion segmentation using object scale-oriented fully convolutional neural networks. Signal Image Video Process 13(3):431–438
32. Ciresan DC, Meier U, Gambardella LM, Schmidhuber J (2010) Deep, big, simple neural nets for handwritten digit recognition. Neural Comput 22(12):3207–3220
33. Wang J, Perez L (2017) The effectiveness of data augmentation in image classification using deep learning. arXiv preprint. arXiv:1712.04621
34. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, London
35. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: proceedings of the 30th IEEE CVPR-2017, pp. 1800–1807. Honolulu, HI, USA
36. Schwarz M, Schulz H, Behnke S (2015) RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. In: proceedings - IEEE International conference robotics automatics, vol. 2015-June, pp. 1329–1335. Washington, USA
37. Sa I, Ge Z, Dayoub F, Upcroft B, Perez T, McCool C (2016) Deepfruits: a fruit detection system using deep neural networks. Sensors (Switzerland) 16(8):1222
38. Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: 31st AAAI-2017. California, USA, San Francisco, pp 4278–4284
39. Howard A.G., Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint. arXiv:1704.04861
40. Dietterich T (2017) Ensemble methods in machine learning, vol. 1857 LNCS (2000)
41. Singh B, Davis L.S. (2018) An analysis of scale invariance in object detection - SNIP. In: proceedings of the IEEE computer society CVPR, pp. 3578–3587. Utah, USA
42. Zhang C, Pan X, Li H, Gardiner A, Sargent I, Hare J, Atkinson PM (2018) A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. J Photogramm Remote Sens 140:133–144
43. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: proceedings of the IEEE CVPR, pp. 770–778. Las Vegas, NV, USA
44. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: proceedings of the IEEE CVPR, vol. 07-12-June-2015, pp. 1–9. Boston, Massachusetts, USA
45. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: international conference on medical image computing and computer-assisted intervention. Springer, Munich, Germany, pp 234–241

46. Al-masni MA, Al-antari MA, Choi MT, Han SM, Kim TS (2018) Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. Comput Meth Prog Bio 162:221–231
47. Lin B.S., Michael K, Kalra S, Tizhoosh H.R. (2018) Skin lesion segmentation: U-Nets versus clustering. In: IEEE SSCI-2017, vol. 2018-Janua, pp. 1–7. Hawaii, USA
48. Alom MZ, Yakopcic C, Hasan M, Taha TM, Asari VK (2019) Recurrent residual U-Net for medical image segmentation. J Med Imaging 6(1):014006
49. Drown DJ, Khoshgoftaar TM, Seliya N (2009) Evolutionary sampling and software quality modeling of high-assurance systems. IEEE Trans Syst Man Cybern Part A Syst Hum 39(5):1097–1107
50. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):1–48
51. Miikkulainen R, Liang J, Meyerson E, Rawal A, Fink D, Francon O, Raju B, Shahrzad H. Navruzyan A, Duffy N, et al. (2019) Evolving deep neural networks. In: artificial intelligence in the age of neural networks and brain computing, pp. 293–312. Elsevier
52. Drown D.J., Khoshgoftaar T.M., Narayanan, R (2007) Using evolutionary sampling to mine imbalanced data. In: Sixth ICMLA-2007, pp. 363–368. IEEE
53. Chan T, Vese L (1999) An active contour model without edges. In: international conference on scale-space theories in computer vision. Springer, Corfu, Greece, pp 141–151
54. Mumford D, Shah J (1989) Optimal approximations by piecewise smooth functions and associated variational problems. Commun Pur Appl Math 42(5):577–685
55. Kowsalya N, Kalyani A, Varsha Shree T.D., Sri Madhava Raja N, Rajinikanth V (2018) Skin-Melanoma evaluation with Tsallis's thresholding and Chan-Vese approach. In: IEEE ICSCA-2018. Pondicherry, India
56. Suzuki K, Wang F, Shen D, Yan P (2011) Machine learning in medical imaging: second international workshop, MLMI 2011, held in conjunction with MICCAI 2011, vol 7009. Springer, Toronto, Canada
57. Setio AAA, Ciompi F, Litjens G, Gerke P, Jacobs C, Van Riel SJ, Wille MMW, Naqibullah M, Sanchez CI, Van Ginneken B (2016) Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE T Med Imaging 35(5):1160–1169
58. Goodfellow I, Bengio Y, Courville A, Bengio Y (2016) Deep learning, vol 1. MIT press, Cambridge
59. Deb K (1996) Genetic algorithms for function optimisation. Genetic Algorithms Soft Comput 8:4–31
60. Radcliffe NJ (1991) Equivalence class analysis of genetic algorithms. Complex Syst 5(2):183–205
61. Herrera F, Lozano M, Verdegay JL (1998) Tackling real-coded genetic algorithms: operators and tools for behavioural analysis. Artif Intell Rev 12(4):265–319
62. Bäck T (1996) Evolutionary algorithms in theory and practice: evolution strategies. Oxford University Press Inc, USA
63. Combalia M, Codella N.C.F., Rotemberg V, Helba B, Vilaplana V, Reiter O, Carrera C, Barreiro A, Halpern A.C., Puig S, Malvehy J (2019) BCN20000: dermoscopic lesions in the wild. arXiv:1908.02288
64. Ballerini L, Fisher R, Aldridge B, Rees J (2013) A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions, vol. 6
65. Tschandl P, Rosendahl C, Kittler H (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5:1–9
66. Gutman D, Codella N.C.F., Celebi E, Helba B, Marchetti M, Mishra N, Halpern A (2016) Skin lesion Analysis toward melanoma detection: a challenge at ISBI-2016, hosted by the international skin imaging collaboration. arxiv:1605.01397
67. Codella N.C.F., Gutman D, Celebi M.E., Helba B, Marchetti M.A., Dusza S.W., Kalloo A, Liopyris K, Mishra N, Kittler H, Halpern A (2018) Skin lesion analysis toward melanoma detection: a challenge at ISBI-2018, hosted by the international skin imaging collaboration. In: proceedings of the international symposium on biomedical imaging, vol. 2018-April, pp. 168–172. Washington, USA
68. Giotis I, Molders N, Land S, Biehl M, Jonkman M, Petkov N (2015) Med-node: a computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Syst Appl 42(19):6578–6585
69. Mendonca T, Ferreira P, Marques J, Marcal A, Rozeira J (2013)Ph2 - a dermoscopic image database for research and benchmarking. In: proceedings of the annual international conference of the IEEE Eng Med Biol Soc, pp. 5437–5440. Osaka, Japan
70. Sun X, Yang J, Sun M, Wang K (2016) A benchmark for automatic visual classification of clinical skin disease images. In: European conference on computer vision. Springer, Amsterdam, The Netherlands, pp 206–222
71. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
72. Rodrigues DDA, Ivo RF, Satapathy SC, Wang S, Hemanth J, Filho PPR (2020) A new approach for classification skin lesion based on transfer learning, deep learning, and IoT system. Pattern Recogn Lett 136:8–15
73. El-Khatib H, Popescu D, Ichim L (2020) Deep learning-based methods for automatic diagnosis of skin lesions. Sensors (Switzerland) 20(6):1753
74. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: proceedings of the 13th AISTATS, pp. 249–256. Sardinia, Italy
75. Dolata P, Mrzygłód M, Reiner J (2017) Double-stream convolutional neural networks for machine vision inspection of natural products. Appl Artif Intell 31(7–8):643–659
76. Yu L, Chen H, Dou Q, Qin J, Heng PA (2017) Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE T Med Imaging 36(4):994–1004
77. Abbasi NR, Shaw HM, Rigel DS, Friedman RJ, McCarthy WH, Osman I, Kopf AW, Polsky D (2004) Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. J Amer Med Assoc 292(22):2771–2776
78. Boughorbel S, Jarray F, El-Anbari M (2017) Optimal classifier for imbalanced data using matthews correlation coefficient metric. PloS One 12(6):e0177678
79. Friedman M (1940) A comparison of alternative tests of significance for the problem of *m* rankings. Ann Math Stat 11(1):86–92
80. Hommel G (1988) A stagewise rejective multiple test procedure based on a modified bonferroni test. Biometrika 75(2):383–386
81. Shaffer JP (1986) Modified sequentially rejective multiple test procedures. J Am Stat Assoc 81(395):826–831
82. Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1(6):80–83
83. Chollet F, et al. (2015) Keras. https://keras.io
84. Abadi M, et al. (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/. Software available from tensorflow.org

85. Gavai N.R., Jakhade Y.A., Tribhuvan S.A., Bhattad R (2017) Mobilenets for flower classification using tensorflow. In: 2017 BID, pp. 154–158
86. Liu X, Jia Z, Hou X, Fu M, Ma L, Sun Q (2019) Real-time marine animal images classification by embedded system based on mobilenet and transfer learning. In: OCEANS 2019 - Marseille, pp. 1–5. https://doi.org/10.1109/OCEANSE.2019.8867190
87. Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33(1–2):1–39
88. Menegola A, Tavares J, Fornaciali M, Li L.T., Avila S, Valle E (2017) RECOD Titans at ISIC Challenge 2017. arXiv:1703.04819

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

6.3. Melanoma Recognition by Fusing Convolutional Blocks and Dynamic Routing between Capsules



Title	Melanoma Recognition by Fusing Convolutional Blocks and Dynamic Routing between Capsules
Authors	E. Pérez, S. Ventura
Journal	Cancers
Volume	13(19)
Year	2021
Editorial	MDPI
DOI	10.3390/cancers13194974
Special issue	Artificial Intelligence in Oncology

IF (JCR 2021)	6.575
Category	Oncology
Position	60/245 (Q1)

cancers

MDPI

Article

Melanoma Recognition by Fusing Convolutional Blocks and Dynamic Routing between Capsules

Eduardo Pérez [1,2] **and Sebastián Ventura** [1,2,3,*]

[1] Knowledge Discovery and Intelligent Systems in Biomedicine Laboratory Maimónides Biomedical Research Institute of Córdoba, 14004 Córdoba, Spain; eduardo.perez@imibic.org
[2] Department of Computer Science and Numerical Analysis, University of Córdoba, 14071 Córdoba, Spain
[3] Department of Information Systems, King Abdulaziz University, Jeddah 21413, Saudi Arabia
[*] Correspondence: sventura@uco.es

Simple Summary: The early treatment of skin cancer can effectively reduce mortality rates. Recently, automatic melanoma diagnosis from skin images has gained attention, which was mainly encouraged by the well-known challenge developed by the International Skin Imaging Collaboration project. The majority of contestant submitted Convolutional Neural Network based solutions. However, this type of model presents disadvantages. As a consequence, Dynamic Routing between Capsules has been proposed to overcome such limitations. The aim of our proposal was to assess the advantages of combining both architectures. An extensive experimental study showed the proposal significantly outperformed state-of-the-art models, achieving 166% higher predictive performance compared to ResNet in non-dermoscopic images. In addition, the pixels activated during prediction were shown, which allows to assess the rationale to give such a conclusion. Finally, more research should be conducted in order to demonstrate the potential of this neural network architecture in other areas.

Abstract: Skin cancer is one of the most common types of cancers in the world, with melanoma being the most lethal form. Automatic melanoma diagnosis from skin images has recently gained attention within the machine learning community, due to the complexity involved. In the past few years, convolutional neural network models have been commonly used to approach this issue. This type of model, however, presents disadvantages that sometimes hamper its application in real-world situations, e.g., the construction of transformation-invariant models and their inability to consider spatial hierarchies between entities within an image. Recently, Dynamic Routing between Capsules architecture (CapsNet) has been proposed to overcome such limitations. This work is aimed at proposing a new architecture which combines convolutional blocks with a customized CapsNet architecture, allowing for the extraction of richer abstract features. This architecture uses high-quality 299 × 299 × 3 skin lesion images, and a hyper-tuning of the main parameters is performed in order to ensure effective learning under limited training data. An extensive experimental study on eleven image datasets was conducted where the proposal significantly outperformed several state-of-the-art models. Finally, predictions made by the model were validated through the application of two modern model-agnostic interpretation tools.

Keywords: melanoma diagnosis; CapsNet; convolutional neural network; interpretation tool

Citation: Pérez, E.; Ventura, S. Melanoma Recognition by Fusing Convolutional Blocks and Dynamic Routing between Capsules. *Cancers* **2021**, *13*, 4974. https://doi.org/10.3390/cancers13194974

Academic Editors: Andreas Stadlbauer, Anke Meyer-Baese and Max Zimmermann

Received: 17 July 2021
Accepted: 29 September 2021
Published: 3 October 2021

1. Introduction

Cutaneous malignant melanoma is on the rise and has the highest mortality rate among the various types of skin cancer [1]. For example, in 2021, it is estimated that 106,110 new cases of melanoma will be diagnosed in the United States, resulting in 7180 deaths (https://www.cancer.org, accessed on 1 June 2021). Surgery is the primary treatment for this type of cancer, but in its more advanced stages, treatment can also include immunotherapy, targeted therapy drugs and radiation to extend survival. Accordingly, the development of modern tools is critical for diagnosing melanoma at an earlier stage, thus easing the decision-making process for dermatologists and reducing invasive treatments for patients, in addition to associated costs. The diagnosis of melanoma is, however, a complex task even for expert dermatologists, mainly because of the complexity, variability and ambiguity of symptoms [2]. Additionally, an extensive variety of morphologies exist even between samples from the same category, which greatly hampers diagnosis. Several studies have shown that the early diagnosis of melanoma can greatly benefit from computational methods [3], demonstrating that such techniques may even outperform dermatologists in terms of diagnosis [4], due to various machine learning techniques and learning of data-driven features for specific tasks [5]. The early proposed methods required the previous extraction of handcrafted features, thus relying on the level of dermatologists' expertise to extract high quality descriptors. This extraction process of informative and discriminative sets of high-level features, however, remains as a complex and costly task that is usually problem dependent [6], and it is noteworthy that sometime is impossible to derive invariant features which are independent of the differences in the input images [7]. On the other hand, there is another type of computational method which can automatically extract and learn high-level features [8], providing a higher robustness to the inter- and intra-class variability present in melanoma images [8,9].

Deep learning models, specifically Convolutional Neural Network (CNN) models, have the capacity of automatically learning high-level features from raw images [8,10,11]. The ImageNet Challenge (ILSVRC) takes place every year since 2010. In 2012 a CNN won the contest for the first time, which increased the popularity of such models for image processing [12]. CNN models learn automatically abstract features and enable the learning for several tasks. For example, Pérez et al. [13] summarized the most popular techniques used in CNN models for diagnosing skin images. Furthermore, this type of deep model can extract sets of patterns ranging from single edges and curves to more complex patterns such as a human face. On the other hand, the main downside of CNN models is that the information regarding spatial relationships between extracted features is lost. For example, CNN models could consider two images to be similar if they share the same objects, even if the location within the image is relevant. However, convolution operation is not translation-invariant.

To overcome the above main limitation of CNN models, a new type of deep learning model, named Dynamic Routing Between Capsules (well-know as CapsNet), was proposed in [14], where the authors designed a method closer to how human vision works. The neurons in this architecture can represent properties of a object such as position, size and texture. Moreover, CapsNet is able to preserve hierarchical spatial relationships, and in theory it could be as effective as any CNN but using fewer samples for training [14]. Niyaz et al. [15] reviewed several deep learning methods for the prediction of different types of cancer. In that time the authors did not find evidence of the application of CapsNet in cancer diagnosis. However, the authors acknowledged CapsNet as a promising model for diagnosing cancer and encouraged its application. Accordingly, CapsNet has been applied in medical image analysis, demonstrating to be really effective for lung cancer screening [16], blood cell image classification [17], and cervical image classification [18], to list a few applications. Finally, CapsNet have been recently applied in skin cancer classification. Cruz et al. [19] used CapsNet to classify skin lesions using images and evaluated their proposal in only one recognized dataset, HAM10000 [20]. However, to our understanding, the proposal has several issues. Firstly, although skin images are usually high-quality (600 × 450 × 3 in HAM10000), the authors resized images to 28 × 28 × 1, losing a considerable amount of pixels and even ignoring colors in the images, which is important for the diagnosis of melanoma [21]. Secondly, the authors highlighted their performance relying mainly on overall precision. However, it is well-known that skin images datasets are unbalanced. Looking closely, the authors achieved only a precision of 28% and 41% in melanoma and basal cell carcinoma categories, respectively, leaving open a big margin of improvement.

Consequently, this work focuses on assessing the effectiveness of a new architecture for the diagnosis of melanoma. The architecture uses high-quality 299 × 299 × 3 skin lesion images and achieves an acceptable performance in both normal and malignant categories. The proposed architecture combines features from convolutional blocks and CapsNet. First, we selected a more sophisticated convolutional computational block, allowing for the extraction of more useful initial features. Second, we replaced the first convolutional block from CapsNet with the above computational block. As a result, we are able to extract more significant features from earlier stages. Next, primary caps extract geometric and color properties present in the images, such as asymmetry, border irregularity, color variegation and the positions of various zones. These have all proven to be very useful attributes to consider when diagnosing melanoma [21]. In this manner, we can maintain the hierarchical spatial relationships of patterns which yields great benefit. To take full advantage of the architecture, we proposes a hyper-tuning of the main parameters to ensure effective training and learning under limited training data. In addition, the architecture applies data augmentation to enhance the diagnosis of melanoma, significantly increasing the validity of the proposal. The new architecture enables the construction of a transformation-invariant model and the detection of spatial hierarchies between entities within an image. As such, it is more suitable for solving certain real-world situations than convolutional models. To evaluate the suitability of the proposal, an extensive experimental study was conducted on eleven public skin image datasets, allowing for a better analysis of the model's effectiveness. The results showed that the proposed approach achieved very promising results and was competitive with respect to state-of-the-art CNN models which have previously been used in the diagnosis of melanoma. Finally, Shapley Additive Explanations method (SHAP (https://github.com/slundberg/shap, accessed on 1 September 2019)) [22] and Local Interpretable Model-agnostic Explanations (LIME) [23] were used to show the most important features and give a prediction with a high confidence level. This work, to the best of our knowledge, is the first attempt to thoroughly assess a new architecture based on convolutional blocks and CapsNet for the automatic recognition of melanoma. The hyper-parameters were specifically tuned for the selected task, achieving significantly better performance compared to the state-of-the-art models.

The rest of this work is arranged as follows: Section 2 briefly presents the state-of-the-art in solving the melanoma diagnosis problem mainly by using CNN models; Section 3 presents the proposed architecture; Section 4 presents the experimental study carried out, showing the results and a discussion of them; finally, some concluding remarks are presented in Section 5.

2. Related Works

CNN models have proven to be a powerful classification method for melanoma diagnosis [8]. This type of models presents a higher suitability compared to classic methods which depend on hand-crafted features. In addition, sophisticated techniques can be applied to even improve the performance of CNN models in the task of melanoma diagnosis, e.g., by applying data augmentation [24] and transfer learning techniques [8].

Data augmentation is a common technique applied to reduce overfitting on CNN models [25]. It is commonly performed by means of applying random transformations on the source images [26]. In addition, this technique can be used to tackle imbalance problems [27,28]. For example, Hossain and Muhammad [29] proposed an emotion recognition system using a CNN approach from emotional Big Data. The models trained with augmented data obtained better performance compared to its non-use. In addition, Esteva et al. [8] applied extensive data augmentation techniques during training; the authors increased the number of images by a factor of 720. Each image was randomly rotated, flipped and cropped. The results achieved a performance comparable to a committee of 21 dermatologists. On the other hand, more advanced techniques such as GANs are being applied to augment data [30]. GANs can augment a dataset by training simultaneously two models, a generator that creates new samples by randomly selecting points from the latent space, and a discriminator that determines whether a sample is a fake or not. Frid-Adar et al. [31] proposed methods based on GANs for generating synthetic medical images; their proposal was evaluated on a limited dataset of high quality liver lesion computed tomography. The results showed that the model increased both sensitivity and specificity by using augmented data.

Transfer learning is a technique widely used to increase performance when the number of training examples is limited [32,33]. This method transfers and reuses knowledge that was learned from a source task, where a lot of data is commonly available, e.g., the ImageNet dataset with more than one million of images. For instance, Esteva et al. [8] transferred the knowledge learned by InceptionV3 on ImageNet and applied it to melanoma diagnosis. Moreover, Nasr-Esfahani et al. [34] applied a pre-trained CNN to distinguishes between melanoma and nevus cases. The results showed that the proposed method is superior in terms of diagnostic accuracy in comparison with the state-of-the-art methods. Finally, Saba et al. [35] proposed an automated approach for skin lesion detection and recognition using Laplacian filtering, lesion boundary extraction and CNN. The results outperformed several existing methods and attained a high accuracy value.

On the other hand, CapsNet represents a completely novel type of deep learning architectures which attempt to overcome the limits and drawbacks of CNN models. Since CapsNet was recently proposed, only a few studies have explored its applications. Zhang et al. [18] applied CapsNet to classify the images of cervical lesions. The results showed better performance compared to other classification methods. Mobiny et al. [16] proposed an improvement on CapsNet that speedup the results compared to the original architecture. After evaluating the performance on computed tomography chest scans, the results showed that CapsNet is a promising alternative to CNN. Zhang et al. [36] combined CapsNet and fully CNN models in image scene classification, such as VGG16 and InceptionV3. The authors achieved better output compared to state-of-the-art methods. However, it is said that the use of a full CNN model could hamper the main aim behind CapsNet, which is the extraction of spatial hierarchies between entities. In addition, the number of trainable parameters significantly increases by combining such architectures. By demonstrating the benefits of CapsNet in medical imaging in this work, we may be encouraging its wider use. Considering the above, it would be interesting to design a deep learning architecture that combines and leverages features from different approaches such as data augmentation, transfer learning, convolutional blocks and CapsNet. After analyzing CapsNet, we strongly believe that specific blocks could be improved while maintaining their behavior. To augment data, it is important to perform a data augmentation both on training and test phases [24]. Next, the proposal for melanoma diagnosis, which follows the mentioned approximation, is described.

3. Materials and Methods

This section firstly describes the related works regarding the automatic diagnosis of melanoma from image data and the well-known state-of-the-art techniques, and then it presents the proposed architecture, which also uses the most proven techniques to date.

3.1. Proposed Architecture for the Diagnosis of Melanoma

Invariance and equivariance are two important concepts in image recognition area. To make a CNN transformation-invariant, a data augmentation of training samples is commonly performed. However, equivariance is a more general concept (invariance is a special case of equivariance) that allows a model detect the rotation or proportion change and adapt itself in a way that the spatial positioning inside an image is not lost [14]. This last requirement motivated the apparition of CapsNet networks.

CapsNet introduced the concept of capsule, where a capsule is a group of neurons or nested set of neural layers, and the state of the neurons inside a capsule can capture the properties of one entity inside an image. A capsule outputs a vector representing the instantiation parameters of a specific type of entity such as an object or a part of a object.

In the other words, the output vector represents the probability of existence. Consequently, similar to the human vision process, these capsules are specialized at handling different types of stimulus and encoding things such as position, size, orientation, deformation, hue, texture, and other spatial information. The output vector can be calculated as

$$v_j = \frac{||s_j||^2}{1 + ||s_j||^2} \frac{s_j}{||s_j||}, \tag{1}$$

where v_j and s_j are the vector output of capsule j and its total output, respectively. The input to a capsule s_j is a weighted sum over the vectors $\hat{u}_{j|i}$ in the layer below and is obtained as

$$s_j = \sum_i c_{ij} \hat{u}_{j|i}, \quad \hat{u} = W_{ij} u_i \tag{2}$$

where W_{ij} is a weight matrix and c_{ij} are coefficients between capsule i and the rest of capsules in the layer above. The coefficients can be calculated as

$$c_{ij} = \frac{exp(b_{ij})}{\sum_k exp(b_{ik})}, \tag{3}$$

where b_{ij} are the log probabilities that capsule i should engage capsule j. Finally, CapsNet uses a separate margin loss, which can be calculated as

$$L_k = T_k \; max(0, m^+ - ||v_k||)^2 + \lambda(1 - T_k) \; max(0, ||v_k|| - m^-)^2, \tag{4}$$

where $T_k = 1$ iff a sample of class k is present and $m^+ = 0.9$ and $m^- = 0.1$. The total loss is the sum of the losses of nevus and melanoma capsules.

The baseline CapsNet architecture is composed by a simplistic Conv2D (256 filters, kernel 9 × 9, stride 1, ReLU activation function [37]), located at the beginning of the network, for extracting primary features which are subsequently passed to Primary and Class Caps layers. However, we hypothesized that CapsNet would attain a better performance if the first convolutional layer is replaced by a more sophisticated convolution-based computational block that was able to extract higher-level features before passing them to capsule layers. By this way, we leverage the benefits from both CNN and CapsNet for a better melanoma diagnosis.

Figure 1 shows the proposed architecture for the diagnosis of melanoma dubbed as MEL-CAP. The proposal was composed as follows: Input (299 × 299 × 3) → Customized convolutional block → Primary Caps (9 × 9, channels 32, capsule 16D) → Class Caps (2 capsules 64D, routing iteration 1). In addition, Table 1 shows a detailed description of every layer. After several phases of an experimental study, the above configuration was the most suitable for diagnosing melanoma. First, Inception architecture was considered given the effectiveness already demonstrated in the diagnosis of melanoma. Inception relies in independent convolutional blocks with filters that are powered the same input, which enables the extraction of more information over the same space. This architecture has been improved through the years, from V1 to V4 [38–40]. The latest updates showed that high performance could be also achieved by using aggressive dimension reductions, which allows to keep low hardware requirements.

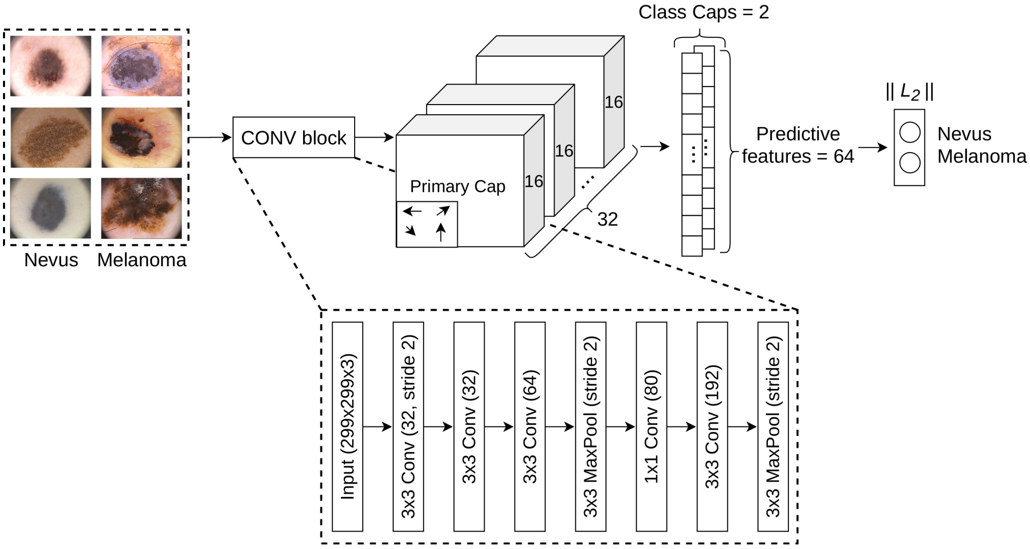


Figure 1. The proposed architecture was designed and hypertuned specifically for the diagnosis of melanoma. Primary caps are able to identify features such as position, size, orientation and deformation. Class caps represent the classes (nevus or melanoma) and resume the predictive features in order to perform the final classification. It was found that each class capsule should select 64 features to perform an accurate prediction. The convolutional block used by the proposed model is shown in the bottom.

Table 1. Proposed network architecture for the diagnosis of melanoma.

Name	Layer	Input	Output
input_1	InputLayer	-	(None, 299, 299, 3)
conv2d_1	Conv2D	(None, 299, 299, 3)	(None, 149, 149, 32)
batch_normalization_1	BatchNormalization	(None, 149, 149, 32)	(None, 149, 149, 32)
activation_1	Activation	(None, 149, 149, 32)	(None, 149, 149, 32)
conv2d_2	Conv2D	(None, 149, 149, 32)	(None, 147, 147, 32)
batch_normalization_2	BatchNormalization	(None, 147, 147, 32)	(None, 147, 147, 32)
activation_2	Activation	(None, 147, 147, 32)	(None, 147, 147, 32)
conv2d_3	Conv2D	(None, 147, 147, 32)	(None, 147, 147, 64)
batch_normalization_3	BatchNormalization	(None, 147, 147, 64)	(None, 147, 147, 64)
activation_3	Activation	(None, 147, 147, 64)	(None, 147, 147, 64)
max_pooling2d_1	MaxPooling2D	(None, 147, 147, 64)	(None, 73, 73, 64)
conv2d_4	Conv2D	(None, 73, 73, 64)	(None, 73, 73, 80)
batch_normalization_4	BatchNormalization	(None, 73, 73, 80)	(None, 73, 73, 80)
activation_4	Activation	(None, 73, 73, 80)	(None, 73, 73, 80)
conv2d_5	Conv2D	(None, 73, 73, 80)	(None, 71, 71, 192)
batch_normalization_5	BatchNormalization	(None, 71, 71, 192)	(None, 71, 71, 192)
activation_5	Activation	(None, 71, 71, 192)	(None, 71, 71, 192)
max_pooling2d_2	MaxPooling2D	(None, 71, 71, 192)	(None, 35, 35, 192)
conv2d_6	Conv2D	(None, 35, 35, 192)	(None, 14, 14, 512)
primary_capsule_reshape	Reshape	(None, 14, 14, 512)	(None, 6272, 16)
primary_capsule_squash	Lambda	(None, 6272, 16)	(None, 6272, 16)
digit_capsule	CapsuleLayer	(None, 6272, 16)	(None, 2, 64)
output_capsule	LengthLayer	(None, 2, 64)	(None, 2)

By this way, we aimed a balance between the computational cost and the extraction of more high-level features before passing them to the capsule layers. Accordingly, Figure 1 also shows the convolutional block that replaced the first convolutional layer of CapsNet. The first block was composed as follows: Conv2D (32 filters, kernel 3 × 3, stride 2) → Conv2D (32 filters, kernel 3 × 3) → Conv2D (64 filters, kernel 3 × 3) → MaxPool2D (3 × 3, stride 2) → Conv2D (80 filters, kernel 1×1) → Conv2D (192 filters, kernel 3×3) → MaxPool2D (3 × 3, stride 2). The use of a convolutional block will not only allow more reduction of the input space, but also focusing on more important features from early stages. On the other hand, capsule layers, will not only learn richer patterns, but also paying more attention on learning their corresponding properties, such as location and orientation.

These abstract features learned by the first block are then passed as input to a convolutional capsule layer, named as primary caps, which is composed by 32 channels of convolutional 8D capsules with a 9 × 9 kernel and stride 2; i.e., in this case, each primary capsule comprises 8 convolutional units. These 8D capsules can identify features such as position, size, orientation, deformation, etc. The last layer, named as class caps, has two capsules 16D that represent the classes (nevus or melanoma), and these capsules receive input from all the capsules in the layer below. Moreover, as proposed in Sabour et al. [14], CapsNet uses a decoder block that influences the learning process, where this decoder intends to reconstruct an original image from the Class Caps layer representation. Finally, it is worth noting that CapsNet implements the routing mechanism mentioned earlier between two consecutive capsule layers (in our example between the layers primary caps and class caps), and this dynamic process can be viewed as a parallel attention mechanism that allows each capsule to attend to some active capsules at the level below and to ignore others [14].

3.2. Datasets

Table 2 shows a summary of the characteristics of the eleven datasets considered in this study, where all the images are labeled as nevus or melanoma. All the datasets were downloaded from The International Skin Imaging Collaboration (https://www.isic-archive.com, accessed on 1 September 2019) (ISIC) repository, except PH2 (https://bit.ly/39YEdmN, accessed on 1 September 2019) and MED-NODE (https://bit.ly/3DkCMvN, accessed on 1 September 2019) datasets. The MED-NODE dataset contains low resolution non-dermoscopic images taken with mobile phones. Nowadays, technological devices enables the collection of an enormous amount of data, which is essential for training models. On the other hand, PH2 dataset comprises high-quality dermoscopic images, where manual segmentation, clinical diagnosis and the identification of several dermoscopic structures were performed by expert dermatologists. The rest of datasets share the common characteristics of dermoscopic images. HAM10000 is the largest dataset in this work, which has been widely used in skin cancer diagnosis, e.g., Miglani and Bhatia [41] achieved 0.95 averaged AUC values for the overall classification. It can be observed that some datasets present a moderate imbalance ratio (ImbR), indicating that the number of nevus samples is several orders of magnitude higher than the number of melanoma samples, and this feature can commonly hamper the learning process of the machine learning models, e.g., MSK-3 and HAM10000. All the 16,601 images were resized to a resolution of $h = 299$, $w = 299$, and $c = 3$, where h is the height, w is the width, and c is the number of channels of an image.

Table 2 also shows other insights about the data. For example, intra-class, inter-class distances and their ratio (DistR) indicate an important degree of similarity between categories. In addition, the silhouette score [42] indicated how much an image shares the same characteristics of its class compared to other classes. The above corroborated that even images from different classes are similar. Finally, in next Section the proposed architecture is evaluated and compared to state-of-the-art CNN models.

Table 2. Skin image datasets used in the experimental study; "ImbR", "IntraC", "InterC" and "DistR" represent the imbalance ratio between the normal and melanoma classes, the average distance between images of the same category, the average distance between images of different categories and the ratio between the two previous metrics, respectively; "Silho" means the silhouette score.

Dataset	Source	# Img	ImbR	IntraC	InterC	DistR	Silho
HAM10000	[20]	7818	6.024	8705	9770	0.891	0.213
ISBI2016	[43]	1273	4.092	10,553	10,992	0.960	0.101
ISBI2017	[44]	2745	4.259	9280	9674	0.959	0.089
MED-NODE	[45]	170	1.429	9029	9513	0.949	0.068
MSK-1	[44]	1088	2.615	11,753	14,068	0.835	0.173
MSK-2	[44]	1522	3.299	9288	9418	0.986	0.062
MSK-3	[44]	225	10.842	8075	8074	1.000	0.112
MSK-4	[44]	943	3.366	6930	7162	0.968	0.065
PH2	[46]	200	4.000	12,688	14,928	0.850	0.210
UDA-1	[43]	557	2.503	11,730	12,243	0.958	0.083
UDA-2	[43]	60	1.609	11,297	11,601	0.974	0.020

4. Analysing the Effectiveness of the Proposal in Melanoma Diagnosis

This section summarizes the experimental study conducted, aiming to analyze the effectiveness of the original CapsNet and our proposal in melanoma diagnosis. First, the experimental protocol and settings used throughout the analysis are described, and finally the experimental results and a discussion of them are presented. Additional material can be found at the available web page (https://www.uco.es/kdis/melanoma-capsnet/, accessed on 25 September 2021).

4.1. Experimental Settings

To test our hypothesis, firstly, three optimization algorithms were used for training the base line CapsNet model: Stochastic Gradient Descend (SGD) [27], Root Mean Square Propagation (RMSProp) [47] and Adaptive Moment Estimation (ADAM) [48]. In this manner, we analyzed what is more convenient for the model: Non-adaptive methods or adaptive gradient descent algorithms. In addition, a binary cross entropy was applied, since the data are comprised of two categories.

Secondly, a hyper-tuning of the two main components of base-line CapsNet was conducted: The dimensions of the primary caps and the activity vector in the class caps. Table 3 shows the hyper-tuning configuration, four dimensions for the primary caps and class caps features were considered. In total 16 combinations were tested with a high computational cost. The best setting obtained is the one that was used in the rest of the experiments.

Table 3. Configuration used in the experimental study.

Parameter	Value
Primary caps	{8, 16, 24, 32}
Class caps features	{16, 32, 48, 64}
Number of epochs	150
Mini-batch size	8
Learning rate (α)	ADAM = 0.001, RMSprop = 0.001, SGD = 0.01
Decay rate first moment average (β_1)	ADAM = 0.9, RMSprop = 0.9
Decay rate second moment average (β_2)	ADAM = 0.999

Thirdly, a data augmentation process was performed both on training and testing phases by means of applying and combining rotation-based, flip-based and crop-based transformations over the original images. The datasets were balanced by creating new images until the number of melanoma images was approximately equal to normal ones. Perez et al. [24] previously demonstrated the benefit of data augmentation process on the melanoma diagnosis problem for constructing more robust CNN models. Consequently, this part of the experimental study aimed at analyzing whether the architectures can be benefited when applying data augmentation as occur with CNN models.

Finally, the performance of our proposal was compared against the following CNN models that have previously been applied in melanoma diagnosis: InceptionV3 [39], DenseNet [49], VGG [50], MobileNet [51], ResNet [52] and EfficientNet [41].

Table 3 shows the basic configuration used for training all the models along the experiments; α, β_1, β_2 were set to the values recommended in the original papers; a batch of size 8 was used due the medium size of the datasets; Xavier method [53] was used to initiate the models; and for non-adaptive optimization methods the learning rate was reduced by a factor of 0.2 when the performance reaches a plateau. Training data were augmented by using random data augmentation techniques, such as rotation, flip and crop transformations. In addition, test data were increased in a different manner. Each test image is augmented 10 times, and the remaining image is linked to the original one. Then, the final prediction is achieved by using a soft-voting strategy.

4.2. Evaluation Process

Regarding the evaluation metrics, Matthews Correlation Coefficient (MCC) and the area under the curve (AUC) values for receiver operator characteristic (ROC) were used to measure the predictive performance of the models, which are commonly applied in Bioinformatics [54,55]. AUC has been recommended in preference to overall accuracy for "single number" evaluation of machine learning algorithms [56]. In addition, MCC and AUC are not biased against the minority class and are commonly used as evaluation metrics to assess the average performance of classifiers on data with imbalanced class distribution [57,58], such as those found in melanoma diagnosis. Both metrics summarize the overall classification performance in a single value for each CNN model. MCC is in the range $[-1, 1]$, where 1 represents a perfect prediction, 0 indicates a performance similar to a random prediction, and -1 an inverse prediction. On the other hand, AUC ranges within $[0, 1]$, where 1 represents a perfect model, 0 the opposite, and 0.5 indicates a random prediction. Nevertheless, it is noteworthy that MCC was used as main metric to measure the predictive performance of the models for melanoma diagnosis in this work, which has been considered before in Alzahrani et al. [59] and Pérez et al. [13].

The results of performing a 3-times 10-fold cross validation were averaged. Finally, significant differences were detected by conducting non-parametric statistical tests with 95% confidence. Friedman's test [60–64] was carried out when a multiple comparison was needed. After that, Hommel's test [65] was applied to detect significant differences with a control algorithm. On the other hand, Wilcoxon Signed-Rank [66] was performed when only two methods were compared.

4.3. Software and Hardware

As for the baseline CapsNet, we used the source code by Xifeng Guo at GitHub (https://bit.ly/3isOBYx, accessed on 1 September 2019). Moreover, the source code to reproduce our work uses Keras v2.2 and TensorFlow v1.12 [67], and can be found at Github (https://bit.ly/3iKOc47, accessed on 25 September 2021). The experimental study was performed in four GPUs Geforce GTX 1080-Ti and four GPUs NVIDIA Geforce RTX 2080-Ti, Intel Core i7-8700K Processor and 64 GB DDR4 RAM.

4.4. Experimental Results

In this section, the most remarkable results of the extensive experimental study are shown; the rest of the experimental study can be consulted at the available web page. Table 4 shows the performance regarding the three diferent optimization algorithms. Results indicated that no significant differences were encountered, showing that CapsNet is not so sensitive regarding the optimization algorithm used; Friedman's test was conducted with two degree of freedom, resulting in a Friedman's statistic equal to 4.136 and p-value equal to 0.126. However, it is worth noting that SGD optimizer occupied the first position of the ranking computed by Friedman's test, meaning that in average CapsNet attained better results when using this optimizer. Consequently, SGD was used as the default optimizer in the rest of experiments.

Table 4. Average MMC values obtained by base-line CapsNet and the three optimization algorithms. The last row shows the average ranking computed by Friedman's test. No significant differences were encountered. The best MCC values and the best ranking were highlighted in bold typeface.

Dataset	ADAM	RMSPROP	SGD
HAM10000	0.065	0.066	**0.242**
ISBI2016	**0.000**	**0.000**	**0.000**
ISBI2017	**0.000**	**0.000**	**0.000**
MED-NODE	0.142	0.182	**0.308**
MSK-1	0.015	0.016	**0.026**
MSK-2	**0.072**	0.024	0.029
MSK-3	**0.000**	**0.000**	**0.000**
MSK-4	0.014	0.014	**0.017**
PH2	0.116	0.159	**0.458**
UDA-1	0.042	0.089	**0.214**
UDA-2	0.132	**0.148**	0.123
Ranking	2.409	2.045	**1.545**

The second part of the experimental study aimed to found the best dimension to primary caps and the number of features for class caps. Table 5 shows the hyper-tuning process on CapsNet architecture. The settings with 16 units in primary caps and 64 features in class caps obtained the first position 8 times, and its closest rival achieved it only 4 times. However, the results indicated that no significant differences were encountered; Friedman's test was conducted with fifteen degree of freedom, resulting in a Friedman's statistic equal to 17.944 and p-value equal to 2.656×10^{-1}. The average ranking showed that the best performance was achieved with 16 units for primary caps and 64 features for each class cap. Furthermore, in this work, the Borda's method [68] was used to compute the average rankings of the individual hyper-parameters. This method is the simplest ranking aggregation method that assigns a score to an element in correspondence to the position in which this element appears in each ranking. Borda's method obtained the ranking for primary caps, being 16, 32, 24 and 8, with 16 as the best and 8 the worst; whereas for class caps was 64, 32, 48 and 16, being 64 the best and 16 the worst. Again, Borda's method obtained that 16 units is the best value for primary caps and 64 for class caps, confirming that a large number of features in class caps means a better predictive performance. Consequently, the best configuration (16 units in primary Caps and 64 features for each class cap) were applied in the rest of the experimental study.

The third part of the experimental study aimed to compare the best configuration for CapsNet model with the proposal. Table 6 shows the average MCC values on test data by using the two models. Firstly, it was observed that the proposed architecture outperformed the base-line CapsNet in all datasets, except in UDA-1 and MSK-3 datasets. In some datasets the differences in performances are remarkable, e.g., in MSK-1 the predictive performance attained by our proposal was 995% higher than CapsNet, in MSK-4 and MSK-2 our proposal was 511% and 485% higher than the base line CapsNet, respectively. In UDA-1 was the only case where our proposal ended 17% behind the base line. In MSK-3 dataset, however, both models presented a performance similar to a classifier making random predictions. In this first comparison, significant differences in performance were encountered, indicating the superiority of our proposal; Wilcoxon' test rejected the null hypothesis with a p-value equal to 3.346×10^{-3}. Secondly, CapsNet and the proposal were also compared by conducting a data augmentation process both on training and test data (as described in Section 4.1). In this case, the results showed that the proposed model outperformed CapsNet in all the datasets applying data augmentation. The differences between our proposal and the base line model were smaller, but still significant; in MSK-3, ISBI2016 and MSK-1 were about 49%, 48% and 39%, respectively. Our proposal achieved 7% better performance than the base line in UDA-1 by applying data augmentation, making our proposal undoubtedly superior in the benchmarks employed. Furthermore, the new architecture obtained a significantly better performance; Wilcoxon's test rejected the null hypothesis with a p-value equal to 1.673×10^{-3}.

The four part of the experimental study focused on comparing the proposal with various CNN models that have previously been used in melanoma diagnosis. We analyzed the MCC and the AUC values in two scenarios: Applying data augmentation and combining data augmentation and transfer learning. Firstly, we analyzed EfficientNet from B0 to B7 in order to select best version. The Friedman's test did not reject the null hypothesis with a p-value equal to 0.841; Friedman's statistic was equal to 3.447 with seven degrees of freedom. However, EfficientNet-B1 obtained the first position in the ranking, followed by EfficientNet-B0. These results are showed at the available web page.

Table 7 shows the average MCC values on test data attained by each model; in this case, a data augmentation process was conducted for all the models. It was observed that the proposal attained the best resultsthe 73% of the time. In addition, the individual percentage improvement of MEL-CAP compared to the state-of-the-arts CNN models are shown. In MSK-3 our architecture achieve a performance 244% and 194% higher than VGG19 and ResNet50, respectively. In UDA-2, the proposal's performance was higher than the rest of the models in at least 17%, and going up to 56% compared with EfficientNet-B1. The biggest differences were located in ISBI2017 and MED-NODE, where our proposal achieved 719% and 511% higher performance than ResNet50. The best performance in the experimental study was achieved in PH2, where all the models obtained above 58.7% of MCC, but even there our proposal was 55% higher than VGG19. The Friedman's test rejected the null hypothesis with a p-value equal to 6.461×10^{-6}; Friedman's statistic was equal to 34.091 with six degrees of freedom. The ranking row of the table shows the average ranking computed by Friedman's test, and this ranking shows that the new model obtained the first position, indicating that this model in average achieved a better performance than the rest of models. Afterwards, the Hommel's post-hoc test was conducted by considering the proposal as the control method, and the results showed the proposal significantly outperformed the rest of the state-of-the-art CNN models.

Table 8 shows the average AUC values on test data attained by each model. The proposed architecture achieved the best average performance in all cases. In MSK-1 our architecture achieve a performance 50% higher than EfficientNet-B1 and ResNet50. In UDA-2, the proposal's performance was higher than the rest of the models in at least 11%. The biggest differences were located in ISBI2017, where our proposal achieved 93% higher performance than ResNet50. The best performance in the experimental study was achieved in PH2, where all the models obtained above 87% of AUC values. The Friedman's test rejected the null hypothesis with a p-value equal to 3.829×10^{-8}; Friedman's statistic was equal to 45.438 with six degrees of freedom. The ranking row of the table shows the average ranking computed by Friedman's test, and this ranking shows that the new model obtained the first position, indicating that this model in average achieved a better performance than the rest of models. Afterwards, the Hommel's post-hoc test was conducted by considering the proposal as the control method, and the results showed the proposal significantly outperformed the rest of the state-of-the-art CNN models.

Table 5. Average MCC values on test data applying the hyper-tuning process on CapsNet architecture. The columns are named with two numbers: First the number of units in primary caps and second the number of features in class caps. The last row shows the average ranking computed by Friedman's test. The best MCC values and the best ranking were highlighted in bold typeface. No significant differences were encountered.

Dataset	8-16	8-32	8-48	8-64	16-16	16-32	16-48	16-64	24-16	24-32	24-48	24-64	32-16	32-32	32-48	32-64
HAM10000	0.242	0.250	0.255	0.257	0.240	0.240	0.246	**0.277**	0.230	0.233	0.239	0.242	0.237	0.244	0.240	0.236
ISBI2016	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
ISBI2017	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
MED-NODE	0.308	0.322	0.376	0.336	0.313	0.342	0.336	**0.412**	0.344	0.361	0.369	0.355	0.346	0.346	0.347	0.347
MSK-1	0.026	0.036	0.030	0.045	0.032	0.042	**0.054**	0.045	0.046	0.046	0.050	0.029	0.051	0.035	0.031	0.028
MSK-2	0.029	0.027	0.029	**0.054**	0.029	0.029	0.037	0.047	0.036	0.036	0.050	0.027	0.050	0.036	0.017	0.026
MSK-3	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
MSK-4	0.017	0.002	0.021	0.029	0.021	0.036	0.021	**0.044**	0.043	0.021	0.017	0.021	0.033	0.033	0.021	0.021
PH2	0.458	0.428	0.412	0.432	0.437	0.455	0.434	**0.516**	0.455	0.455	0.455	0.455	0.455	0.451	0.451	0.437
UDA-1	0.214	0.210	0.225	0.225	0.218	0.201	0.236	**0.377**	0.200	0.207	0.187	0.199	0.199	0.206	0.201	0.191
UDA-2	0.123	0.187	0.122	0.096	0.168	0.251	0.140	0.236	0.201	0.201	0.201	**0.264**	0.201	0.207	0.201	**0.264**
Ranking	10.182	10.455	8.955	7.909	10.045	7.909	7.909	**4.091**	8.227	7.955	8.182	8.727	7.545	7.727	9.818	10.364

Table 6. Average MCC values on test data by using the hyper-tuned CapsNet architecture; CAP represents the base-line CapsNet and MEL-CAP represents the proposal; the last three columns show a comparison between the same architecture but only applying data augmentation both in train and test data. Moreover, we showed the percent improvement of the proposal compared to base-line CapsNet, e.g., MEL-CAP achieved 78% higher performance than CAP in HAM10000. The best MCC values were highlighted in bold typeface. The labels "Inf" represent those cases where a base-line model obtained an average MCC value equal to zero.

Dataset	CAP	MEL-CAP	%	CAP	MEL-CAP	%
HAM10000	0.277	**0.493**	78.0	0.698	**0.896**	28.4
ISBI2016	0.000	**0.234**	Inf	0.499	**0.740**	48.3
ISBI2017	0.000	**0.184**	Inf	0.640	**0.819**	28.0
MED-NODE	0.412	**0.485**	17.7	0.608	**0.671**	10.4
MSK-1	0.045	**0.493**	995.6	0.575	**0.801**	39.3
MSK-2	0.047	**0.275**	485.1	0.600	**0.694**	15.7
MSK-3	**0.000**	**0.000**	0.0	0.525	**0.782**	49.0
MSK-4	0.044	**0.269**	511.4	0.694	**0.752**	8.4
PH2	0.516	**0.644**	24.8	0.849	**0.909**	7.1
UDA-1	**0.377**	0.310	−17.8	0.503	**0.542**	7.8
UDA-2	0.236	**0.559**	136.9	0.531	**0.601**	13.2
p-value	3.346×10^{-3}			1.673×10^{-3}		

Table 7. Average MCC values on test data obtained by the proposal and state-of-the-art CNN models when applying data augmentation. The best MCC value attained in each dataset is highlighted in bold typeface. The percentage means the difference between MEL-CAP versus the other CNN models, e.g., MEL-CAP attained 20% percent of improvement compared to EfficientNet-B1 in HAM10000 dataset. In addition, it is shown the overall average and the ranking computed by Friedman's test. Last row shows multiple comparisons between the new architecture (control model) and state-of-the-art CNN models through Hommel's post-hoc test.

Dataset	InceptionV3	DenseNet201	VGG19	MobileNet	ResNet50	EfficientNet-B1	MEL-CAP
HAM10000	0.873 (+3%)	0.753 (+19%)	0.649 (+38%)	0.760 (+18%)	0.510 (76%)	0.746 (20%)	**0.896**
ISBI2016	0.655 (+13%)	0.656 (+13%)	0.511 (+45%)	0.575 (+29%)	0.403 (+84%)	**0.798** (−7%)	0.740
ISBI2017	0.749 (+9%)	0.715 (+15%)	0.575 (+42%)	0.744 (+10%)	0.100 (+719%)	0.800 (2%)	**0.819**
MED-NODE	0.618 (+9%)	0.514 (+31%)	0.540 (+24%)	0.660 (+2%)	0.100 (+571%)	0.502 (34%)	**0.671**
MSK-1	0.754 (+6%)	0.792 (+1%)	0.610 (+31%)	0.785 (+2%)	0.466 (+72%)	0.481 (67%)	**0.801**
MSK-2	0.518 (+34%)	0.631 (+10%)	0.428 (+62%)	0.531 (+31%)	0.358 (+94%)	0.635 (9%)	**0.694**
MSK-3	0.565 (+38%)	0.588 (+33%)	0.227 (+244%)	0.532 (+47%)	0.266 (+194%)	**0.903** (−13%)	0.782
MSK-4	0.693 (+9%)	0.696 (+8%)	0.467 (+61%)	0.596 (+26%)	0.370 (+103%)	0.573 (31%)	**0.752**
PH2	0.840 (+8%)	0.778 (+17%)	0.587 (+55%)	0.902 (+1%)	0.819 (+11%)	0.862 (5%)	**0.909**
UDA-1	0.489 (+11%)	0.501 (+8%)	**0.555** (−2%)	0.535 (+1%)	0.430 (+26%)	0.430 (12%)	0.542
UDA-2	0.471 (+28%)	0.408 (+47%)	0.412 (+46%)	0.403 (+49%)	0.514 (+17%)	0.386 (56%)	**0.601**
Ranking	3.636	3.818	5.273	3.727	6.273	4.000	**1.273**
p-values	1.029×10^{-2}	1.029×10^{-2}	7.044×10^{-5}	1.029×10^{-2}	3.417×10^{-7}	1.027×10^{-2}	-

Table 8. Average AUC values on test data obtained by the proposal and state-of-the-art CNN models when applying data augmentation. The best AUC values and the best ranking were highlighted in bold typeface.

Dataset	InceptionV3	DenseNet201	VGG19	MobileNet	ResNet50	EfficientNet-B1	MEL-CAP
HAM10000	0.876 (+2%)	0.875 (+2%)	0.778 (+15%)	0.875 (+2%)	0.869 (+3%)	0.862 (+4%)	**0.895**
ISBI2016	0.849 (+6%)	0.842 (+6%)	0.771 (+16%)	0.855 (+5%)	0.646 (+39%)	0.878 (+2%)	**0.896**
ISBI2017	0.862 (+6%)	0.878 (+4%)	0.820 (+12%)	0.858 (+7%)	0.474 (+93%)	0.863 (+6%)	**0.915**
MED-NODE	0.767 (+7%)	0.780 (+5%)	0.652 (+26%)	0.765 (+7%)	0.420 (+96%)	0.678 (+21%)	**0.822**
MSK-1	0.821 (+8%)	0.862 (+3%)	0.776 (+14%)	0.844 (+5%)	0.590 (+50%)	0.591 (+50%)	**0.886**
MSK-2	0.831 (+5%)	0.841 (+3%)	0.727 (+20%)	0.829 (+5%)	0.582 (+49%)	0.703 (+24%)	**0.869**
MSK-3	0.921 (+3%)	0.920 (+3%)	0.896 (+6%)	0.880 (+8%)	0.758 (+25%)	0.927 (+2%)	**0.946**
MSK-4	0.845 (+10%)	0.886 (+5%)	0.842 (+10%)	0.882 (+5%)	0.687 (+35%)	0.764 (+21%)	**0.926**
PH2	0.882 (+6%)	0.909 (+3%)	0.905 (+3%)	0.922 (+2%)	0.871 (+7%)	0.880 (+6%)	**0.936**
UDA-1	0.770 (+6%)	0.780 (+4%)	0.723 (+13%)	0.809 (+1%)	0.649 (+25%)	0.673 (+21%)	**0.814**
UDA-2	0.638 (+18%)	0.630 (+20%)	0.650 (+16%)	0.678 (+12%)	0.637 (+19%)	0.679 (+11%)	**0.756**
Ranking	3.727	3.227	5.273	3.500	6.727	4.545	**1.000**
p-values	9.206×10^{-3}	1.561×10^{-2}	1.754×10^{-5}	1.329×10^{-2}	3.028×10^{-9}	4.744×10^{-4}	-

Table 9 shows the average MCC on test data attained by each model when applying data augmentation and transfer learning. The proposed architecture achieved the best average performance in all cases, except in ISBI2016, ISBI2017 and PH2. The biggest differences were located in MED-NODE, where our proposal achieved 166% higher performance than ResNet50. In UDA-2, the proposal's performance was higher than the rest of the models in at least 15%. The best performance in the experimental study was achieved in PH2, where all the models obtained above 84% of MCC values. The Friedman's test rejected the null hypothesis with a *p*-value equal to 3.189×10^{-8}; Friedman's statistic was equal to 45.838 with six degrees of freedom. The ranking row of the table shows the average ranking computed by Friedman's test, and this ranking shows that the new model obtained the first position, indicating that this model in average achieved a better performance than the rest of models. Afterwards, the Hommel's post-hoc test was conducted by considering the proposal as the control method, and the results showed the proposal significantly outperformed the rest of the state-of-the-art CNN models, except MobileNet and DenseNet.

Table 9. Average MCC values on test data obtained by the proposal and state-of-the-art CNN models when applying data augmentation and transfer learning. The best MCC values and the best ranking were highlighted in bold typeface.

Dataset	InceptionV3	DenseNet201	VGG19	MobileNet	ResNet50	EfficientNet-B1	MEL-CAP
HAM10000	0.940 (+4%)	0.954 (+3%)	0.601 (+63%)	0.945 (+3%)	0.870 (+12%)	0.809 (+21%)	**0.978**
ISBI2016	0.802 (+10%)	0.850 (+3%)	0.625 (+41%)	0.850 (+3%)	0.385 (+128%)	**0.945** (−7%)	0.879
ISBI2017	0.829 (+8%)	0.854 (+5%)	0.738 (+22%)	0.875 (+3%)	0.414 (+117%)	**0.929** (−3%)	0.899
MED-NODE	0.732 (+5%)	0.698 (+10%)	0.486 (+58%)	0.741 (+4%)	0.289 (+166%)	0.568 (+35%)	**0.768**
MSK-1	0.868 (+3%)	0.880 (+1%)	0.708 (+26%)	0.886 (+0%)	0.350 (+154%)	0.598 (+49%)	**0.890**
MSK-2	0.805 (+9%)	0.830 (+5%)	0.561 (+56%)	0.860 (+2%)	0.350 (+150%)	0.738 (+18%)	**0.874**
MSK-3	0.959 (+4%)	**1.000**	0.911 (+10%)	**1.000**	0.606 (+65%)	**1.000**	**1.000**
MSK-4	0.844 (+8%)	0.864 (+5%)	0.825 (+10%)	0.890 (+2%)	0.482 (+89%)	0.710 (+28%)	**0.910**
PH2	0.963 (+3%)	0.960 (+3%)	0.923 (+8%)	0.963 (+3%)	0.836 (+19%)	**1.000** (−1%)	0.993
UDA-1	0.720 (+13%)	0.764 (+7%)	0.585 (+40%)	0.781 (+5%)	0.463 (+76%)	0.632 (+29%)	**0.817**
UDA-2	0.413 (+60%)	0.522 (+27%)	0.477 (+39%)	0.577 (+15%)	0.485 (+36%)	0.452 (+46%)	**0.661**
Ranking	4.409	3.273	5.818	2.500	6.545	4.045	**1.409**
p-values	4.506×10^{-3}	8.610×10^{-2}	8.482×10^{-6}	2.363×10^{-1}	1.475×10^{-7}	1.263×10^{-2}	-

Table 10 shows the average AUC values on test data attained by each model when applying data augmentation and transfer learning. The proposed architecture achieved the best average performance in all cases, except in ISBI2016 and PH2. The biggest differences were located in MSK-1, where our proposal achieved 40% higher performance than ResNet50. The best performance in the experimental study was achieved in PH2, where all the models obtained above 93% of AUC. The Friedman's test rejected the null hypothesis with a *p*-value equal to 3.518×10^{-8}; Friedman's statistic was equal to 45.623 with six degrees of freedom. The ranking row of the table shows the average ranking computed by Friedman's test, and this ranking shows that the new model obtained the first position, indicating that this model in average achieved a better performance than the rest of models. Afterwards, the Hommel's post-hoc test was conducted by considering the proposal as the control method, and the results showed the proposal significantly outperformed the rest of the state-of-the-art CNN models, except MobileNet and DenseNet.

Table 10. Average top AUC values on test data obtained by the proposal and state-of-the-art CNN models when applying data augmentation and transfer learning. The best AUC values and the best ranking were highlighted in bold typeface.

Dataset	InceptionV3	DenseNet201	VGG19	MobileNet	ResNet50	EfficientNet-B1	MEL-CAP
HAM10000	0.990 (+0%)	0.993 (+0%)	0.841 (+18%)	0.992 (+0%)	0.960 (+4%)	0.913 (+9%)	**0.994**
ISBI2016	0.936 (+3%)	0.946 (+2%)	0.843 (+14%)	0.950 (+1%)	0.728 (+32%)	**0.990** (−3%)	0.961
ISBI2017	0.957 (+2%)	0.962 (+2%)	0.916 (+7%)	0.963 (+1%)	0.748 (+31%)	0.960 (+2%)	**0.977**
MED-NODE	0.868 (+3%)	0.844 (+5%)	0.734 (+21%)	0.864 (+3%)	0.640 (+39%)	0.763 (+17%)	**0.890**
MSK-1	0.941 (+2%)	0.957 (+1%)	0.864 (+12%)	0.950 (+1%)	0.687 (+40%)	0.697 (+38%)	**0.964**
MSK-2	0.935 (+3%)	0.943 (+2%)	0.802 (+20%)	0.945 (+2%)	0.702 (+37%)	0.843 (+14%)	**0.961**
MSK-3	0.995 (+1%)	**1.000**	0.986 (+1%)	**1.000**	0.850 (+18%)	**1.000**	**1.000**
MSK-4	0.954 (+3%)	0.960 (+2%)	0.927 (+6%)	0.967 (+1%)	0.778 (+26%)	0.858 (+14%)	**0.979**
PH2	0.991 (+1%)	0.991 (+1%)	0.981 (+2%)	0.991 (+1%)	0.934 (+7%)	**1.000** (0%)	0.998
UDA-1	0.875 (+4%)	0.891 (+2%)	0.808 (+13%)	0.899 (+1%)	0.743 (+23%)	0.811 (+12%)	**0.912**
UDA-2	0.700 (+21%)	0.738 (+15%)	0.738 (+15%)	0.779 (+9%)	0.742 (+14%)	0.739 (+15%)	**0.847**
Ranking	4.364	3.273	5.864	2.591	6.455	4.136	**1.318**
p-values	3.783×10^{-3}	6.769×10^{-2}	4.015×10^{-6}	1.671×10^{-1}	1.475×10^{-7}	6.652×10^{-3}	-

4.5. Explanation of the Predictions

The results showed that our proposal was effective for solving the melanoma diagnosis problem. In addition, we encourage the explanations of the individual predictions and as a result, in this work the areas where the proposal paid more attention were showed. To do this, SHAP and LIME were applied. The first one determines how much a pixel contributes to the diagnosis in comparison with the overall result. On the other hand, LIME consists in identifying an interpretable model over the interpretable representation that is locally faithful to the classifier. As a result, the super-pixels with positive weight towards the predicted class were highlighted, as they give intuition as to why the model would think that class should be selected. Figure 2 shows how cluster of pixels were activated in the proposed architecture. The first image was classified as nevus, meanwhile the another one was classified as melanoma. Red pixels represent positive Shapley values that increase the probability of being melanoma, while blue pixels represent negative Shapley values that increase the probability of being non-melanoma. On the other hand, the weight of each superpixel is showed by LIME, where blue pixels mean those that most support the prediction and red ones the lower support. Both analyses corroborated that the proposed model focused in the lesion itself.

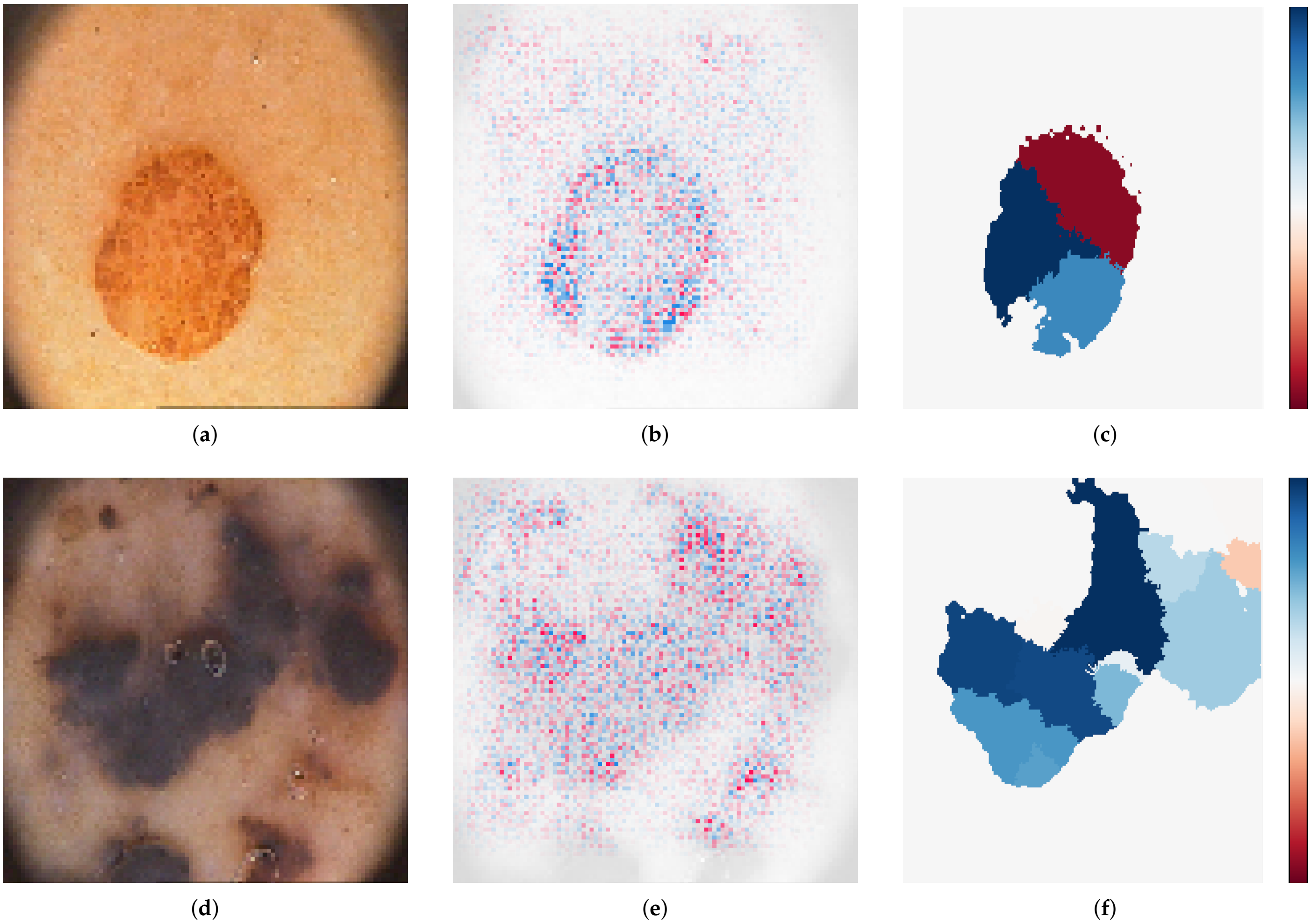


Figure 2. Model-agnostic interpretation tools. (**a**) Nevus, (**b**) SHAP, (**c**) LIME, (**d**) Melanoma, (**e**) SHAP, (**f**) LIME.

5. Conclusions

In this work, a novel neural network architecture for diagnosing melanoma has been proposed, allowing the early extraction of richer abstract features before passing them to deeper layers. The use of CapsNet combined with convolutional blocks allowed a better learning of the representations. By this way, better predictive features could be extracted, thus facilitating the learning of better abstract and discriminative features for melanoma diagnosis. The proposed architecture is flexible regarding the design of its blocks. Consequently, custom networks could easily be designed, for example by employing another convolutional block with a simpler or more complex internal structure. Moreover, the predictive features from CapsNet could be used to feed other well-known models, such as Support Vector Machine, which has proven to achieve high performance [69]. The results corroborated that data augmentation and transfer learning are suitable techniques to improve the proposal and all studied CNN models, overcoming common issues in melanoma diagnosis, such as small datasets and imbalance data. Finally, the proposed model significantly outperformed state-of-the-art CNN models that haven previously been applied for solving melanoma diagnosis problem, confirming the potential that possess this novel neural network architecture.

The research on CapsNet is still at early stage and, therefore, few application on real-world problems can be found so far. Consequently, more research and extensive experimental study should be conducted in order to demonstrate and confirm the full potential of this neural network architecture. As future works, we will also design ensemble learning techniques for a better application in small and medium problems. Finally, we encourage further development of the research line that combines the proposal and other CNN models for a better melanoma diagnosis.

Author Contributions: Conceptualization, E.P.; methodology, E.P. and S.V.; software, E.P.; validation, E.P.; formal analysis, E.P.; investigation, E.P.; resources, E.P. and S.V.; data curation, E.P.; writing—original draft preparation, E.P.; writing—review and editing, E.P. and S.V.; visualization, E.P.; supervision, S.V.; project administration, S.V.; funding acquisition, S.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Spanish Ministry of Science and Innovation, the University of Cordoba, and the European Regional Development Fund, projects PID2020-115832GB-I00 and UCO-FEDER 18 REF.1263116 MOD. It was also funded by the i-PFIS contract no. IFI17/00015 granted by Health Institute Carlos III of Spain.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data sources used in this manuscript are publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

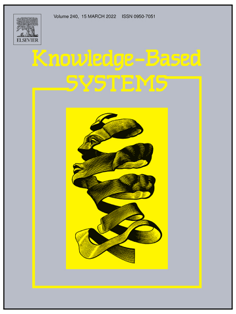
References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2019. *CA Cancer J. Clin.* **2019**, *69*, 7–34. [CrossRef] [PubMed]
2. Geller, A.C.; Swetter, S.M.; Brooks, K.; Demierre, M.F.; Yaroch, A.L. Screening, early detection, and trends for melanoma: Current status (2000–2006) and future directions. *J. Am. Acad. Dermatol.* **2007**, *57*, 555–572. [CrossRef] [PubMed]
3. Lee, H.D.; Mendes, A.I.; Spolaôr, N.; Oliva, J.T.; Sabino Parmezan, A.R.; Wu, F.C.; Fonseca-Pinto, R. Dermoscopic assisted diagnosis in melanoma: Reviewing results, optimizing methodologies and quantifying empirical guidelines. *Knowl.-Based Syst.* **2018**, *158*, 9–24. [CrossRef]
4. Haenssle, H.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Ben Hadj Hassen, A.; Thomas, L.; Enk, A.; Uhlmann, L. Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **2018**, *29*, 1836–1842. [CrossRef] [PubMed]
5. Rastgoo, M.; Lemaître, G.; Morel, O.; Massich, J.; Garcia, R.; Mériaudeau, F.; Marzani, F.; Sidibé, D. Classification of melanoma lesions using sparse coded features and random forests. In *Medical Imaging 2016: Computer-Aided Diagnosis*; International Society for Optics and Photonics: Bellingham, WA, USA, 2016; Volume 9785.
6. Jin, L.; Gao, S.; Li, Z.; Tang, J. Hand-crafted features or machine learnt features? together they improve RGB-D object recognition. In Proceedings of the IEEE International Symposium on Multimedia (ISM-2014), Miami, FL, USA, 14–16 December 2015; pp. 311–319.
7. Liu, X.; Wang, X.; Matwin, S. Interpretable Deep Convolutional Neural Networks via Meta-learning. In Proceedings of the International Joint Conference on Neural Networks, Rio de Janeiro, Brazil, 8–13 July 2018.
8. Esteva, A.; Kuprel, B.; Novoa, R.; Ko, J.; Swetter, S.; Blau, H.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
9. Abbes, W.; Sellami, D. High-level features for automatic skin lesions neural network based classification. In Proceedings of the 2nd International Image Processing, Applications and Systems Conference, Singapore, 4–6 August 2017.
10. Munir, K.; Elahi, H.; Ayub, A.; Frezza, F.; Rizzi, A. Cancer diagnosis using deep learning: A bibliographic review. *Cancers* **2019**, *11*, 1235. [CrossRef] [PubMed]
11. Alyafeai, Z.; Ghouti, L. A fully-automated deep learning pipeline for cervical cancer classification. *Expert Syst. Appl.* **2020**, *141*, 112951. [CrossRef]
12. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *2*, 1097–1105. [CrossRef]
13. Pérez, E.; Reyes, O.; Ventura, S. Convolutional neural networks for the automatic diagnosis of melanoma: An extensive experimental study. *Med. Image Anal.* **2021**, *67*, 101858. [CrossRef]
14. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic Routing Between Capsules. *arXiv* **2017**, arXiv:1710.09829.
15. Niyaz, U.; Sambyal, A.S. Advances in deep learning techniques for medical image analysis. In Proceedings of the Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Himachal Pradesh, India, 20–22 December 2018; pp. 271–277.
16. Mobiny, A.; Van Nguyen, H. Fast capsnet for lung cancer screening. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018; pp. 741–749.
17. Zhang, X.; Zhao, S. Blood Cell Image Classification Based on Image Segmentation Preprocessing and CapsNet Network Model. *J. Med. Imaging Health Inform.* **2019**, *9*, 159–166. [CrossRef]

18. Zhang, X.; Zhao, S.G. Cervical image classification based on image segmentation preprocessing and a CapsNet network model. *Int. J. Imaging Syst. Technol.* **2019**, *29*, 19–28. [CrossRef]
19. Cruz, M.V.; Namburu, A.; Chakkaravarthy, S.; Pittendreigh, M.; Satapathy, S.C. Skin cancer classification using convolutional capsule network (CapsNet). *J. Sci. Ind. Res.* **2020**, *79*, 994–1001.
20. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **2018**, *5*, 1–9. [CrossRef]
21. Abbasi, N.R.; Shaw, H.M.; Rigel, D.S.; Friedman, R.J.; McCarthy, W.H.; Osman, I.; Kopf, A.W.; Polsky, D. Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria. *J. Am. Med. Assoc.* **2004**, *292*, 2771–2776. [CrossRef] [PubMed]
22. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
23. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
24. Perez, F.; Vasconcelos, C.; Avila, S.; Valle, E. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*; Springer: Granada, Spain, 2018; pp. 303–311.
25. Wang, J.; Liu, Q.; Xie, H.; Yang, Z.; Zhou, H. Boosted efficientnet: Detection of lymph node metastases in breast cancer using convolutional neural networks. *Cancers* **2021**, *13*, 1–14.
26. Wang, J.; Perez, L. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv* **2017**, arXiv:1712.04621.
27. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
28. Menegola, A.; Tavares, J.; Fornaciali, M.; Li, L.T.; Avila, S.; Valle, E. RECOD Titans at ISIC Challenge 2017. *arXiv* **2017**, arXiv:1703.04819.
29. Hossain, M.S.; Muhammad, G. Emotion recognition using deep learning approach from audio–visual emotional big data. *Inf. Fusion* **2019**, *49*, 69–78. [CrossRef]
30. Liu, W.; Luo, Z.; Li, S. Improving deep ensemble vehicle classification by using selected adversarial samples. *Knowl.-Based Syst.* **2018**, *160*, 167–175. [CrossRef]
31. Frid-Adar, M.; Diamant, I.; Klang, E.; Amitai, M.; Goldberger, J.; Greenspan, H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **2018**, *321*, 321–331. [CrossRef]
32. Wang, S.; Yang, D.; Rong, R.; Zhan, X.; Fujimoto, J.; Liu, H.; Minna, J.; Wistuba, I.; Xie, Y.; Xiao, G. Artificial intelligence in lung cancer pathology image analysis. *Cancers* **2019**, *11*, 1673. [CrossRef] [PubMed]
33. Kaymak, R.; Kaymak, C.; Ucar, A. Skin lesion segmentation using fully convolutional networks: A comparative experimental study. *Expert Syst. Appl.* **2020**, *161*, 113742. [CrossRef]
34. Nasr-Esfahani, E.; Samavi, S.; Karimi, N.; Soroushmehr, S.; Jafari, M.; Ward, K.; Najarian, K. Melanoma detection by analysis of clinical images using convolutional neural network. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Orlando, FL, USA, 16–20 August 2016; pp. 1373–1376.
35. Saba, T.; Khan, M.A.; Rehman, A.; Marie-Sainte, S.L. Region Extraction and Classification of Skin Cancer: A Heterogeneous framework of Deep CNN Features Fusion and Reduction. *J. Med. Syst.* **2019**, *43*, 1–19. [CrossRef] [PubMed]
36. Zhang, W.; Tang, P.; Zhao, L. Remote Sensing Image Scene Classification Using CNN-CapsNet. *Remote Sens.* **2019**, *11*, 494. [CrossRef]
37. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. *J. Mach. Learn. Res.* **2011**, *15*, 315–323.
38. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
39. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
40. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-2017), San Francisco, CA, USA, 4–9 February 2017; pp. 4278–4284.
41. Miglani, V.; Bhatia, M. Skin lesion classification: A transfer learning approach using efficientnets. *Adv. Intell. Syst. Comput.* **2021**, *1141*, 315–324._29. [CrossRef]
42. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
43. Gutman, D.; Codella, N.C.F.; Celebi, E.; Helba, B.; Marchetti, M.; Mishra, N.; Halpern, A. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *arXiv* **2016**, arXiv:1605.01397.

44. Codella, N.C.F.; Gutman, D.; Celebi, M.E.; Helba, B.; Marchetti, M.A.; Dusza, S.W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC-2018). In Proceedings of the International Symposium on Biomedical Imaging, Washington, DC, USA, 4–7 April 2018; pp. 168–172.
45. Giotis, I.; Molders, N.; Land, S.; Biehl, M.; Jonkman, M.; Petkov, N. MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Syst. Appl.* **2015**, *42*, 6578–6585. [CrossRef]
46. Mendonca, T.; Ferreira, P.; Marques, J.; Marcal, A.; Rozeira, J. PH2 - A dermoscopic image database for research and benchmarking. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Osaka, Japan, 3–7 July 2013; pp. 5437–5440.
47. Hinton, G.; Srivastava, N.; Swersky, K. Rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude. Neural Networks for Machine Learning, Coursera Lecture 6e. 2012. Available online: https://bit.ly/3ooAQxN (accessed on 1 September 2019).
48. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
49. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Densely connected convolutional networks. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2017), Honolulu, HI, USA, 21–26 July 2016.
50. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
51. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
52. Han, S.S.; Kim, M.S.; Lim, W.; Park, G.H.; Park, I.; Chang, S.E. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *J. Investig. Dermatol.* **2018**, *138*, 1529–1538. [CrossRef] [PubMed]
53. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
54. Faraggi, D.; Reiser, B. Estimation of the area under the ROC curve. *Stat. Med.* **2002**, *21*, 3093–3106. [CrossRef]
55. Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* **2017**, *12*, e0177678. [CrossRef] [PubMed]
56. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
57. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C.A.F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412–424. [CrossRef] [PubMed]
58. Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Handling imbalanced datasets: A review. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *30*, 25–36.
59. Alzahrani, S.; Al-Bander, B.; Al-Nuaimy, W. A Comprehensive Evaluation and Benchmarking of Convolutional Neural Networks for Melanoma Diagnosis. *Cancers* **2021**, *13*, 4494. [CrossRef] [PubMed]
60. Friedman, M. A comparison of alternative tests of significance for the problem of *m* rankings. *Ann. Math. Stat.* **1940**, *11*, 86–92. [CrossRef]
61. Jiang, Y.; Cukic, B.; Menzies, T. Can data transformation help in the detection of fault-prone modules? In Proceedings of the 2008 Workshop on Defects in Large Software Systems, Seattle, WA, USA, 20–20 July 2008; pp. 16–20.
62. Jerez, J.M.; Molina, I.; García-Laencina, P.J.; Alba, E.; Ribelles, N.; Martín, M.; Franco, L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif. Intell. Med.* **2010**, *50*, 105–115. [CrossRef] [PubMed]
63. Liang, G.; Zhu, X.; Zhang, C. An empirical study of bagging predictors for imbalanced data with different levels of class distribution. In Proceedings of the Australasian Joint Conference on Artificial Intelligence, Perth, Australia, 5–8 December 2011; pp. 213–222.
64. Umigai, N.; Murakami, K.; Ulit, M.; Antonio, L.; Shirotori, M.; Morikawa, H.; Nakano, T. The pharmacokinetic profile of crocetin in healthy adult human volunteers after a single oral administration. *Phytomedicine* **2011**, *18*, 575–578. [CrossRef] [PubMed]
65. Hommel, G. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika* **1988**, *75*, 383–386. [CrossRef]
66. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics* **1945**, *1*, 80–83. [CrossRef]
67. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: Tensorflow.org (accessed on 1 September 2019).
68. Dwork, C.; Kumar, R.; Naor, M.; Sivakumar, D. Rank aggregation methods for the web. In Proceedings of the 10th International Conference on World Wide Web, Hong Kong, China, 1–5 May 2001; pp. 613–622.
69. Khan, M.A.; Javed, M.Y.; Sharif, M.; Saba, T.; Rehman, A. Multi-model deep neural network based features extraction and optimal selection approach for skin lesion classification. In Proceedings of the International Conference on Computer and Information Sciences (ICCIS-2019), Aljouf, Saudi Arabia, 10–11 April 2019.

6.4. Diagnosing melanoma via active learning and convolutional models



Title	A framework to build accurate Convolutional Neural Network models for melanoma diagnosis
Authors	E. Pérez, and S. Ventura
Journal	Knowledge-Based Systems
Status	Submitted, 2nd revision round
Year	2021
Editorial	Elsevier

IF (JCR 2021)	8.139
Category	Computer Science - Artificial Intelligence
Position	24/145 (Q1)

Highlights

A framework to build accurate Convolutional Neural Network models for melanoma diagnosis

Eduardo Pérez, Sebastián Ventura

- A more stable training process for melanoma diagnosis, which is inspired on Active Learning and Convolutional Neural Networks, is proposed.
- The batch-based query strategy setting proposed in this work enables a more faster training process by learning about the complexity of the data.
- The proposed method was assessed by analyzing the informativeness value of each image and the predictive performance of the models.
- The proposal outperforms five state-of-the-art Convolutional Neural Networks models using 16 skin lesion datasets.

A framework to build accurate Convolutional Neural Network models for melanoma diagnosis

Eduardo Pérez[b,a], Sebastián Ventura[a,b,c,*]

[a] *Andalusian Research Institute in Data Science and Computacional Intelligence, DaSCI, University of Cordoba, Cordoba, 14071, Andalusia, Spain*
[b] *Maimonides Biomedical Research Institute of Cordoba, IMIBIC, University of Cordoba, Cordoba, 14071, Andalusia, Spain*
[c] *Department of Information Systems, King Abdulaziz University, Saudi Arabia Kingdom*

Abstract

In the past few years, Convolutional Neural Networks have achieved performance levels similar to those achieved by dermatologists. However, the diagnosis of melanoma remains a challenging task, mainly due to the high levels of inter and intra-class variability present in images of moles. Aimed at developing new methods for an effective melanoma diagnosis, a new framework is proposed. The training process is expertly guided by an active learning approach where the architectures implicitly learn about the complexity of individual images through query strategies, which allows us to adjust the training process and achieve better performance. In addition, we propose a batch-based query strategy that enables a more stable and faster training process. Also, the framework leverages segmentation, data augmentation and transfer learning to enhance melanoma diagnosis. The framework is composed by several specialized blocks, which allow us to measure how the diagnosis is improved after each step. In this sense, blocks could be customized and do not depend on specific models. An extensive experimental study was conducted on 16 skin image datasets, where five state-of-the-art models were significantly outperformed. This study corroborated that new active learning query strategies can be employed to effectively train neural networks architectures for the diagnosis of melanoma, achieving 182% better predictive performance in Xception, and an overall 11% and 20% better predictive performance in dermoscopic and non-dermoscopic images, respec-

[*] Corresponding author: sventura@uco.es

Preprint submitted to Knowledge-Based Systems *June 28, 2022*

tively. In addition, the informativeness value of each image is shown, which leads to identify the hardest images for the predictive models. Finally, the proposal required 2% of the total training time, and needed 61% less training epochs.

Keywords: melanoma diagnosis, convolutional neural networks, active learning, query strategy, data augmentation, transfer learning, segmentation

1. Introduction

Melanoma is the most dangerous type of skin cancer. It can appear anywhere, but it is most often found on areas of the body that are exposed to the sun. For example, in the United States, 106,110 new cases of invasive melanoma and 7,180 deaths are expected in 2021 [1]. This requires further research and the development of improved diagnostic methods. However, despite the progress made during recent years, the diagnosis of melanoma is still a challenge due to the uncertainty and complexity of its symptoms [2]. In addition, differences between samples of the same illness can be found, making the diagnosis even more difficult. Several methods have been applied in the diagnosis of melanoma, including the clinical procedure known as Asymmetry, Border, Color, Diameter and Evolution (ABCDE) which helps dermatologists differentiate between skin lesions [3]. However, despite their experience, dermatologists rarely achieve test sensitivities greater than 80% [4]. If doubts remain, a biopsy is performed.

On the other hand, computational methods such as Convolutional Neural Network (CNN) has been demonstrated that may match or even surpass diagnoses made by dermatologists [5]. CNN models automatically learn high-level abstract features from raw images and do not need previously extracted characteristics [6]. CNN models began being used for image processing after AlexNet [7] won the well-known *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) in 2012. Since 2016, *The International Skin Imaging Collaboration*[1] (ISIC) has hosted a challenge in which more than 180 teams have participated. Competitors solve challenges related to lesion segmentation, detection of clinical diagnostic patterns and lesion classification. In addition, several more advanced pipelines have been proposed. For example,

[1] https://www.isic-archive.com

Kaur et al. [8] proposed a segmentation framework by using a custom architecture based on atrous/dilated convolutions, which enabled the proposal to extract lesions in less time. The authors claim higher performance than the best ones of the ISIC contests regarding segmentation. However, bear in mind that the previous proposal needs a training step, which is avoided in the current segmentation block that will be presented in the next steps of our proposal. Also, Basak et al. [9] researched on various scaled feature maps for computing the final segmentation mask. The authors built a custom network based on Res2Net [10], which achieved better performance than state-of-the-art methods in three well-known skin datasets. On the other hand, Lafraxo et al. [11] proposed a framework for detecting melanoma in dermoscopic images. Different techniques were used in order to avoid over-fitting, such as dropout and data augmentation, which achieved an acceptable performance on three publicly available datasets. However, non-dermoscopic images were tested at discretion, with only one dataset evaluated. Nowadays, the number of non-dermoscopic images is increasing sharply, which is noteworthy.

Most of the above solutions are based on CNN models. However, despite having proven to be successful in solving several complex problems [12], CNN models still present several issues that hamper their correct application in the diagnosis of melanoma. It is noteworthy that most of the existing public skin images datasets encompass only a few hundred or thousands of images, and at the same time CNN models fit a wide variety of non-linear data points. The above combination can lead to overfitting on datasets with small numbers of training examples per class, therefore attaining a poor generalization capacity, e.g. as is the case in the UDA-2 dataset [13] with only 60 images. In addition, we analyzed a large number of datasets and found that they present large inter-class similarities and intra-class variances. CNN models are sensitive to the above characteristics and also to variations in viewpoints, changes in lighting conditions, occlusions and background clutter [14], which can be also found in skin cancer images. On top of that, although the majority of melanoma datasets consist of dermoscopic images verified and labeled by expert dermatologists, there is a growing tendency to collect images taken with common digital cameras, which requires models to accurately predict both types of images. New capable models can boost the development of modern tools for a less expensive melanoma diagnosis, thus reducing invasive treatments and the required economical resources [15]. Finally, when taking a picture of the lesion, there is not a right position to do it. Bear in mind that CNN models are invariant to small transformations on the input, but not invariant on rotation, color and lighting [16]; invariance to transformation is an important concept in image recognition and applies perfectly in skin images. If you take the input and transform it, the representation you end up with is the same as the representation of the original, which is what happens when you analyze skin images. Due to the large morphologic variation of moles, it is important to allow a model to detect the rotation or proportion changes and obtain the same abstract representation, thus attaining a more effective melanoma diagnosis.

Taking into account the above limitations, the most proven and widely used techniques to improve the performance of CNN models for melanoma diagnosis are transfer learning and data augmentation [17]. For example, Esteva et al. [5] applied a trained InceptionV3 model and augmented the dataset by a large factor of 720×. The authors achieved a performance on par with 21 experts, proving the suitability of CNN models and these techniques for helping dermatologists to classify skin cancer. Although the above research is one of the most important for melanoma diagnosis, we believe that the computational resources to carry out their proposal are expensive. On the other hand, despite the advantages that other techniques offer for improving CNN models, such as ensemble learning, generative adversarial networks [18], multi-task learning methods [19, 20], and segmentation, it is worth highlighting some limitations in its application. Training ensemble models requires evaluating a high number of possibilities that increase the computational resources needed, such as the type of CNN models to combine, and how to combine the learned abstract features and individual predictions for each member. As for generative adversarial networks, these architectures require a complex training process where the generative and discriminative models need to be in balance, and commonly fail to converge [21]. As for multi-task learning, the lack of datasets that contain heterogeneous information is the main limitation when applying this paradigm for diagnosing melanoma, such as images and meta-data. Regarding segmentation methods, we believe one limitation is the lack of datasets with binary segmentation masks indicating the lesion areas, which are difficult, time-consuming and expensive to obtain [22]. Also, graphics processing unit (GPU) based segmentation methods are expensive to train, which is a handicap if we are already using GPU to train CNN models to make predictions, increasing the computational resources required.

Finally, active learning has recently been applied to select a small subset of images from where the model learns [23]. However, CNN models usually need tens of thousands to hundred thousands samples to achieve the best performance. For example, Shi et al. [24] used only a selection of the instances and did not achieve better performance in any evaluation metric compared to other proposals. However, the amount of available data in skin datasets is usually small, hampering the default application of the active learning approach in melanoma diagnosis. As a result, our proposal explores the entire pool of training data and organizes the mini-batches from where the CNN model learns. In this manner, after each epoch, the active learning query strategy extracts the informativeness value from each sample and decides when the CNN model analyzes that sample. We believe it is important to improve the training process itself, rather than incorporating more techniques disproportionately. In most cases, researchers rely on their expertise to select which techniques to apply, and there is no specific pattern that will produce a model with a high level of reliability. Furthermore, most studies do not follow a standard experimental study and only include a limited number of datasets.

As a consequence of the above, in this work, a framework is presented, aiming to train CNN models and improve their performance in melanoma diagnosis. Commonly, a CNN model is trained following a mini-batch gradient descent approach, which consists in randomly splitting the training set in small batches in order to calculate the model error. However, in this work a novel training process is proposed following an Active Learning (AL) approach [25, 26], where AL determines which images go in every mini-batch. Nevertheless, the framework is not limited to melanoma diagnosis, and it could be applied to other real-world problems in the future. This is thanks to using query strategies, which calculate how informative each sample is for a CNN model. The aim is to prove that removing the randomness during training, and replacing it by a training process expertly guided by an AL approach, the architectures will implicitly learn about the complexity of individual images. Such complexities allow us to adjust the training process and achieve better performance in a lower number of iterations compared to random mini-batch sampling. In this manner it is possible to apply state-of-the-art AL query strategies or even implement new ones according to the context. In addition, we propose a new query strategy that samples both “easy” and “hard” instances, which guarantees a more balanced and controlled training process. Then, the models learn about the complexities of individual images and change its training sequence in the next epoch, achieving a high performance even with limited training data. In addition, the framework applies an extension of the Chan-Vese segmentation algorithm for skin lesion images [27, 28], which do not need a training phase or previously segmented images, basic data augmentation techniques and transfer learning. It is well known that preprocessing the input data can improve the quality of any biomedical data analysis. In addition, data augmentation, which is performed to reduce overfitting and to obtain transformation-invariant models [29], and transfer learning from pre-trained ImageNet allow to alleviate the requirement for a large number of training data. These proposals allow us to train CNN models in a more consistent way. Evaluation protocols and details are critical to support significant empirical findings.

Experiments corroborated that all CNN models trained using the proposed framework significantly outperformed state-of-the-art methods. The segmentation block enabled a maximum gain of 11% in diagnostic performance when using NASNet as classifier. On the other hand, the proposed query strategy achieved 182% and 72% better predictive performance in Xception and Inception, respectively. Furthermore, we provided an analysis focused on the AL component of the framework, understanding how the query strategies sample the instances and corroborating that query strategies are able to control the mini-batch training process. Finally, the best predictive performance is found in significantly fewer epochs.

The rest of this work is organized as follows: Section 2 briefly presents the state-of-the-art in solving melanoma diagnosis problem mainly by using CNN models; Section 3 describes the proposed framework; the experimental design is shown in Section 4; the analysis and discussion of the results are portrayed in Section 5; finally, some highlights are presented in Section 6.

2. Related works

CNN models can automatically learn a set of abstract features from raw data and achieve a high performance without the need for extracting handcrafted features [30]. However, there is a need to further improve performance to help dermatologists in decision making. Some authors have consistently achieved a significant high performance applying sophisticated techniques over CNN models to attain a better melanoma diagnosis, e.g. applying segmentation [31], basic data augmentation [17] and transfer learning [5], which we applied in our proposal and explained below.

Skin lesion segmentation is a highly complex challenge for melanoma diagnosis. For instance, from 2016 to 2018 there was a special task in ISIC challenges related specifically to lesion segmentation. In ISIC-2016, considering the top three average results, there was a slight improvement in the use of segmentation methods compared to its non-use, confirming its importance. Ronneberger et al. [32] designed a CNN model dubbed U-Net for biomedical image segmentation which relies on the use of data augmentation to use the available labeled images more efficiently. The U-Net architecture achieved an acceptable performance on different biomedical segmentation applications such as neuronal structures in electron microscopic recordings and skin lesion segmentation [33]. On the other hand, Alom et al. [34] proposed a recurrent residual U-Net model (dubbed R2U-Net). The model was tested on blood vessel segmentation in retinal images, skin cancer segmentation, and lung lesion segmentation. However, both U-Net and R2U-Net require a costly training process using GPU and more importantly, a prior knowledge from training data already segmented by expert dermatologists.

Another widely used technique is data augmentation, which is commonly applied to build a low-variance model. It is performed applying random transformations on the original images. For example, Esteva et al. [5] applied data augmentation on the largest melanoma dataset reported to the date, each image was randomly rotated, cropped and flipped. On the other hand, Lenc and Vedaldi [16] evaluated CNN models by applying data augmentation and found that deeper models benefited the most when this technique was applied.

In addition, transfer learning is a method which transfer and reuse a knowledge that was learned from a source task [35], where a lot of data is commonly available, e.g. the ImageNet dataset with more than one million of images. Several authors have shown the usefulness of transfer learning in melanoma diagnosis. Esteva et al. [5] used a pre-trained InceptionV3 on ImageNet and applied it in melanoma diagnosis. Khan et al. [36] applied pre-trained ResNet50 and ResNet101, extracting the best learned features that after were used to build a Support Vector Machine (SVM) model.

Finally, AL is used in this work to lead the training process and avoid the randomness involved in the mini-batches. AL is a subfield of machine learning and its hypothesis is that, *"if a learning algorithm is allowed to select the data from which it learns, it can achieve a better performance with fewer labeled data"*. AL has been applied in several computer-aided scenarios. For example, Zliobaite et al. [37] explored AL in data stream settings and proposed three AL strategies for streaming data that explicitly handle concept drift. The method focuses on selecting as few labeled instances as possible for learning an accurate predictive model. On the other hand, Lin et al. [38] studied multiclass imbalance problems and proposed a sampling method for multilayer perceptrons. The method selected which data should be used to train the model. Furthermore, AL has been applied to improve the performance in CNN models. Sener and Savarese [23] applied AL to train CNN models more efficiently in the Canadian Institute For Advanced Research datasets, named CIFAR-10 and CIFAR-100 datasets. The authors selected a subset of images from which the model learned and the method significantly outperformed existing approaches by a large margin. However, the amount of available data in the 81% of skin datasets is 60× smaller compared to the above datasets and also the resolution in skin cancer images is greater, e.g. CIFAR-10 has 60,000 32x32 images compared to PH2 with 200 765x572 images. This hampers the default application of the AL approach in the diagnosis of melanoma. Considering the above, we hypothesized that if CNN models were actively trained by smartly selecting which images go in each mini-batch, it could lead to achieve a better melanoma diagnosis. In essence, AL asks queries about the instances and these results are reflected in the next training epochs. The proposal explores the entire pool of training data and organizes the mini-batches from where the CNN model learns. In this manner, after each epoch the AL query strategy extracts the informativeness value from each sample and decides when the CNN model analyzes that sample. The framework applies the most proven techniques, improving the performance of a CNN model after each step. Next, following the mentioned approximation, a new framework for melanoma diagnosis is described.

3. A framework to build accurate Convolutional Neural Network models

Let us say $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is a dataset of n images, where x_i represents an image and y_i is the class. Let φ be a model that follows a CNN architecture, which learns the representations from the feature space and yields a prediction. Once the prediction for a given training image is computed, the loss obtained by applying φ on the ith training image ($\mathcal{L}(i)$) is computed by means of a binary cross entropy

$$\mathcal{L}(i) = -y_i \cdot \log(\hat{y}_i) - (1 - y_i) \cdot \log(1 - \hat{y}_i) \tag{1}$$

where $\hat{y}_i$ is the probability predicted by φ indicating that the ith training image is melanoma or nevus, and y_i is the actual class of the image; in our case, this probability can be computed by applying a sigmoid activation function on an output layer. Consequently, the goal is to find the set of parameters θ that controls φ such that the following empirical risk is minimized along the n training samples and φ,

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(i) \tag{2}$$

To solve this optimization problem, any gradient descent algorithm that iteratively minimizes the prediction errors on the training samples can be used, since the first derivative of the loss function used is well-defined. The derivative of $J(\theta)$ with respect to the set of weight parameters of φ (θ) is computed as

$$\frac{\partial J}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \mathcal{L}(i)}{\partial \theta} \tag{3}$$

where $\mathcal{L}(i)$ is the loss function of φ.

There are three main ways of training a CNN model: stochastic, batch, and mini-batch. Stochastic gradient descent calculates the error and updates the model for each example in the training dataset. However, updating the model so frequently is more computationally expensive. On the other hand, batch gradient descent calculates the error for each example, but it only updates the model after all training examples have been evaluated. However, the more stable error gradient may result in premature convergence of the model to a less optimal set of parameters. Also, commonly, batch gradient descent implementations require the entire training dataset in memory and available to the algorithm, which is very computationally expensive. Finally, mini-batch gradient descent splits the training dataset into small batches that are used to calculate model error. The above method has several advantages, such as that the model update frequency is higher than batch gradient descent, which allows a more robust convergence, avoiding local minima; batch-based updates provide a computationally more efficient process than stochastic gradient descent; and the split in small batches allows the efficiency of not having all training data in memory. All images in the batch are processed in parallel using GPU memory, significantly increasing the training speed; also, small batches can serve as a regularizing effect [39]. Therefore, in this work the CNN models were trained using mini-batch gradient descent. Commonly, during each epoch each mini-batch is formed by images taken randomly from the training set. However, we propose a different approach.

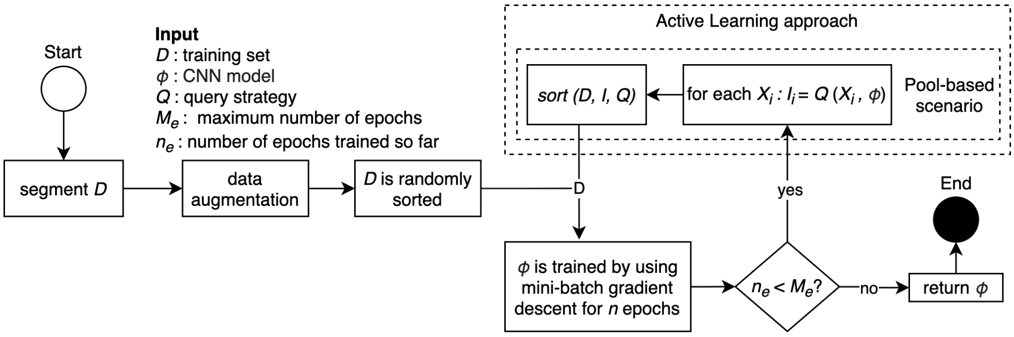


Figure 1: Steps of the proposed Active Learning framework.

We hypothesize that if a CNN model is allowed to select in which order it trains over the images it can achieve a better performance in fewer iterations. We are aware that CNN models commonly require a large collection of data in order to build an accurate model. In this context AL is applied not in a traditional way, which is building φ with the minimum number of training samples. Instead of that, AL provides us the sequence of images from which the model trains.

Figure 1 shows how the proposed framework is integrated within the training process of the model φ. Firstly, the images are segmented following our own extension of the Chan-Vese algorithm [27]. The above algorithm is designed to segment objects without clearly defined boundaries and is based on techniques of curve evolution, Mumford-Shah [40] functional for segmentation and level sets. The first step is to apply the above method on the input images in order to obtain a mask with positive and negative values. Then, our custom procedure was introduced, which experimentally showed to be effective for skin cancer image analysis. We select the p_c of the positive pixels within the centre, which is motivated by Suzuki et al. [41], where the authors encouraged the use of images centered on the lesion in order to perform a more accurate analysis in medical imaging. Then, all positive clusters of pixels that intersect p_c are merged, obtaining a new segmentation mask. Finally, the segmented image is obtained after applying the new mask on the original image. Bear in mind that p_c could be tuned depending on the datasets.

Secondly, data augmentation is applied in the segmented images, increasing the number of images, reducing overfitting and attaining a high generalization capacity. Thirdly, the resulting images are sorted randomly. Fourthly,

φ is trained for N epochs using mini-batch gradient descent following the given order. Fifthly, φ computes $\hat{y}_i$ and the query strategy (Q) calculates the informativeness values over all D following a *pool-based* active learning scenario [26]. Sixthly, D is sorted according the informativeness values and Q; the query strategy decides whether it is ascending or descending. Then, steps from IV to VI are repeated until a stop condition is satisfied. Finally, a model φ is returned. Next, it is explained in detail how to calculate the informativeness values by using AL in this particular scenario.

Let us say Q and $I = \{I_1, I_2, \ldots, I_n\}$ are the query strategy and the set of values associated with dataset D, respectively, where I_i represents how informative the sample X_i is for φ. Generally speaking, every N epochs, Q analyzes D and updates I as follows: $I_i = Q(X_i, \varphi)$. Then D is sorted according I. After that, each mini-batch is assembled considering the values of I. For instance, let us say that each mini-batch is composed by b samples. The first mini-batch will be composed by the b most informative and the last mini-batch by the b least informative samples. Scenarios and query strategies are two main components of AL [26]. The scenario in AL controls how is the interaction between the oracle, the CNN model and the query strategy. There are two main scenarios, stream-based selective sampling and pool-based active learning. Stream-based selective sampling assumes that evaluating an image is free (or inexpensive), so it can first be sampled from the actual distribution, and then the CNN model can decide whether or not to request its evaluation, usually using a threshold. This scenario has the disadvantage that by directly selecting an image it may be ignoring whether any of the following is more informative for the CNN model. On the other hand, pool-based active learning assumes that there is a pool of samples available. Images are greedily queried, according to an informativeness measure used to evaluate all images in the pool. This scenario has been applied before in other tasks such as text classification and image classification [42]. The above scenario is the one that best suited our problem, because we need to keep the entire training set sorted.

The selected AL scenario requires evaluating the informativeness of the pool of samples. There have been many proposed query strategies. However, in this framework AL is used in a particular manner, so only the most feasible query strategies in providing a measure of uncertainty about each sample are considered. The simplest and most commonly used query strategy is Uncertainty Sampling (US), which also has few variants, such as *Relevance Sampling* (RS) and *Least Confident Sampling* (LC) [43, 44]. Nevertheless, other more complex query strategies are compatible with the framework and could be applied, such as *Query-by-committee* (QBC) [45] and *Density-Weighted methods* (ID) [46].

In LC, a CNN model queries the instances where there is less certainty about how to label them, and the opposite for RS. For our problem, the informativeness values can be calculated as

$$x^*_{LC} = {}^{\arg\min} x \ \hat{y}^*_i \tag{4a}$$

$$x^*_{RS} = {}^{\arg\max} x \ \hat{y}^*_i \tag{4b}$$

where $\hat{y}^*_i$ means the probability of the most likely class labeling and x^*_{LC} and x^*_{RS} represent the most informative samples for each query strategy, respectively. For example, considering a binary classification scenario, x^*_{LC} will be the sample with $\hat{y}_i$ closest to 0.5 and x^*_{RS} will be the one with highest $\hat{y}_i$. In other words, for LC query strategy the first images to train are those harder to predict for the CNN model. On the other hand, RS put first the images from those the model feels more positive about its prediction.

In QBC, a committee $C = \varphi_1, \varphi_2, ..., \varphi_n$ of models is considered, e.g. MobileNet, DenseNet and Xception. The models train on the same training set, but represent different hypotheses. Each member is allowed to vote on the query images. Commonly, the most informative image is considered to be the one which the committee members most disagree. The aim is to minimize the version space, which represents the set of hypotheses that are consistent with the training data. For measuring the level of disagreement, there are several approaches, two of them are *vote entropy* (VE) and *Kullback-Leibler divergence* (KL), and can be calculated as

$$x^*_{VE} = {}^{\arg\max} x - \sum_i \frac{V(y_i)}{C} \log(\frac{V(y_i)}{C}) \tag{5}$$

$$x^*_{KL} = {}^{\arg\max} x \frac{1}{C} \sum_{c=1}^{C} D(P_{\varphi_j} || P_C) \tag{6}$$

$$D(P_{\varphi_j} || P_C) = \sum_i P(y_i|x;\varphi_j) \log \frac{P(y_i|x;\varphi_j)}{P(y_i|x;C)} \tag{7}$$

where y_i is the predicted label, $V(y_i)$ is the number of votes for that label, φ_j represents the model j in the committee of models C, and finally $P(y_i|x;C) =$

$\frac{1}{C}\sum_{c=1}^{C} P(y_i|x;\varphi_j)$ is the agreement probability that y_i is the right label. However, the training of this type of model commonly requires the assessment of a high number of possibilities that hamper the process, such as the number and type of CNN models to combine. US and QBC are prone to querying outliers, which has been the motivation for creating other more advanced query strategies that also consider representativeness between samples. In ID, the aim is to find which sample is the most uncertainty for φ, but also is representative of the training set D; this query strategy can be calculated as

$$x^*_{ID} = {}^{\arg\max} x \omega_A(x) \times \left(\frac{1}{D} \sum_{d=1}^{D} \mathrm{sim}(x, x^d) \right)^{\beta} \qquad (8)$$

where $\omega_A(x)$ means the informativeness value of x according to another baseline query strategy A, such as US and QBC. The second term calculates the average similarity of the current sample to all other instances in the training set.

The idea of sampling both "easy" and "hard" samples has previously shown potential advantages compared to the baseline random method. Both approaches follow different paths when training, but reduce variance, achieve different local minima, and are better compared to random mini-batch sampling. In this work we also proposed a new query strategy named Uncertainty Mixed Sampling (UMix), combining the aforementioned strategies in a simple manner. During odd epochs the strategy will behave as Eq. 4a, firstly analyzing complex samples for the model, which have previously achieved high performance in a faster way [47]. On the other hand, at even epochs the proposal will change to Eq. 4b, which is inspired by how children learn, from samples of increasing difficulty [48]. In this manner, the training process is implicitly balanced and controlled by the learning needs of φ. An analysis about how the query strategies and the models sample the instances is shown in Section 5. In this way, we are able to prove our initial hypothesis where the models learn about the complexities of individual instances and automatically change their behavior after each epoch.

4. Experiments

This section describes the experimental design carried out in this work. First, the datasets and the experimental protocol are portrayed, and then, the evaluation process and the software and hardware are presented.

4.1. Image datasets for melanoma diagnosis

To test the proposal, dermoscopic and non-dermoscopic images were obtained from several reputable sources [6]. Table 1 shows a summary of the benchmark datasets. Only the images labeled as melanoma and nevus were considered, being in total 36,703 images. Most datasets present high imbalance ratio (ImbR), up to 10 in the case of MSK-3, commonly hampering the learning process. The intra-class (IntraC) and inter-class (InterC) metrics show the average distances between images belonging to different classes, as well as between images belonging to the same class. Both metrics were computed using the Euclidean function distance; each image i was represented as a vector. Then, the ratio (DistR) between these metrics showed that both distances are similar, which commonly indicates a high degree of overlapping between classes. Finally, the silhouette score (Silho) [49] was calculated, representing how similar an image is to its own cluster compared to other clusters. The results indicated that images were not well matched to their own cluster, and even samples belonging to different clusters are close in the feature space.

4.2. Experimental settings

In the first phase of the experimental study, the proposed segmentation algorithm was applied on PH2, DERM-LIB, ISIC-2016 and ISIC-2017, where these datasets were manually segmented by expert dermatologists and are commonly used as benchmarks due its high quality [50, 51]. The aim was to find how much the method is close to expert segmentation. Also, we compared the segmentation algorithm to U-Net and R2U-Net. Bear in mind that the above CNN architectures need prior training. In order to obtain the segmentation mask of an image X, the CNN architectures were trained during 150 epochs with 10% of images as validation set and the rest as training set. The best model obtained during validation was applied on the image X and a segmentation mask was obtained.

In the second phase of the experimental study the framework was evaluated with the following state-of-the-art CNN models that have previously been used in melanoma diagnosis [6]: InceptionV3, DenseNet, MobileNet, Xception and NASNetMobile. In this phase each model was compared itself as follows: (I) without any technique, (II) applying transfer learning and data augmentation in train and test, (III) the same as (II) but using the segmented images, and finally, (IV) the same as (III) but training with the proposed *active learning mini-batch sampling* method.

Table 1: Summary of the benchmark datasets. The datasets are grouped in dermoscopic and non-dermoscopic.

Dataset	Img	ImbR	IntraC	InterC	DistR	Silho
BCN20000	17,393	2.848	9,014	10,107	0.892	0.153
DERM-LIB	407	4.355	7,171	9,163	0.783	0.270
DERM7PT-D	827	2.282	15,971	16,866	0.947	0.087
HAM10000	7,818	6.024	8,705	9,770	0.891	0.213
ISBI2016	1,273	4.092	10,553	10,992	0.960	0.101
ISBI2017	2,745	4.259	9,280	9,674	0.959	0.089
MSK-1	1,088	2.615	11,753	14,068	0.835	0.173
MSK-2	1,522	3.299	9,288	9,418	0.986	0.062
MSK-3	225	10.842	8,075	8,074	1.000	0.112
MSK-4	943	3.366	6,930	7,162	0.968	0.065
PH2	200	4.000	12,688	14,928	0.850	0.210
UDA-1	557	2.503	11,730	12,243	0.958	0.083
UDA-2	60	1.609	11,297	11,601	0.974	0.020
DERM7PT-C	827	2.282	15,442	16,318	0.946	0.086
MED-NODE	170	1.429	9,029	9,513	0.949	0.068
SDC-198	648	4.735	14,054	14,840	0.947	0.116

The configuration listed in Table 2 was used to train all the models for any of the experiments as following: the learning rate (α) was equal to 0.01 and it was reduced by a factor of 0.2 if an improvement in predictive performance was not observed during 10 epochs; the weights of the networks were initialized using Xavier method [52]; a batch of size 8 was used due the medium size of the used datasets; and the models were trained along 150 epochs. *Stochastic Gradient Descent* (SGD) [53] was used for training the models, SGD is one of the most used optimizers for training CNN models and despite its simplicity, it performs well across a variety of applications and has been successfully applied for training networks in melanoma diagnosis [17, 6]. Training CNN models is an arduous task, therefore we selected *Relevance Sampling*, *Least Confident Sampling* and the proposal UMix as query strategies. The aim of this work is to corroborate that training CNN models powered by an AL approach can lead to better performance, rather than finding the best possible configuration. Regarding the tuning of the hyper-parameters of SGD, it is noteworthy that finding the optimal set of the hyper-parameter values is a task that commonly requires expensive and

Table 2: Basic configuration used.

Parameter	Value
p_c	40%
Rotations	[1°,270°]
Flip	vertical and horizontal
Translations in X and Y	[-30%,30%]
Crop	[10%,30%]
Number of epochs	150
Mini-batch size	8
Learning rate (α)	SGD=0.01
Factor for decreasing α	0.2
Data augmentation factor on test	10×
Final prediction	soft-voting strategy
Query strategies	Relevance sampling Eq. 4b, Least confident Eq. 4a, Uncertainty Mixed Sampling

arduous work due to the many possible combinations [54]. In this work, a tuning process was not carried out and so the results could not be conferred to an over-adjustment. The datasets utilized in this work correspond to binary classification problems, so the cost function used for training the models was defined as the average of the binary cross entropy along all training samples. Data augmentation technique was applied to tackle the imbalance problem by applying and combining basic transformations. After splitting a dataset in training and test sets, training data were balanced by creating new images until the number of melanoma images was equal to normal ones, and the generated training images were considered as independent from the original ones. On the other hand, test data were expanded by randomly augmenting each test image at least 10 times, but the generated images remained related to the original ones. As a result, the classes' probabilities for any test image and its related set of images were averaged using a soft-voting strategy to yield a final prediction. Finally, we made available the source code[2] in order to extend the present work.

4.3. Evaluation process

Regarding the evaluation process of the CNN models, a 3-times 10-fold cross validation process was conducted on each dataset and the results were

[2] https://github.com/eperezp1990/melanoma-al

averaged across all fold executions. As can be seen in Table 1, the imbalance ratio in skin image datasets is large (from 1.6× to 10.8×), e.g. in HAM10000 there are 6× more normal samples compared to malignant. As a result, in each fold execution, *Matthews Correlation Coefficient* (MCC) was used to measure the predictive performance of the models. MCC is widely used in Bioinformatics as a performance metric [55, 56, 57, 6, 58], and it is specially designed to analyze the predictive performance on unbalanced data, even if the classes are of very different sizes. MCC gives a good summary of a confusion matrix, and it is computed as

$$MCC = \frac{t_p \times t_n - f_p \times f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}} \tag{9}$$

where t_p, t_n, f_p, and f_n are the number of true positives, true negative, false positives, and false negatives, respectively. MCC is in range $[-1, 1]$, where 1 represents a perfect prediction, 0 indicates the model is performing similarly to random prediction, and -1 an inverse prediction. Non-parametric statistical tests were used to detect whether there was any significant difference in predictive performance. Friedman's test was conducted in cases where a multiple comparison was carried out, and afterward, Hommel's post-hoc test [59] was employed to perform a multiple comparison with a control algorithm. On the other hand, Wilcoxon Signed-Rank test [60] was performed in those cases where only two individual methods were compared. All hypothesis testing were conducted at 95% confidence.

4.4. Software and hardware

The experimental study was executed in Ubuntu 18.04, Intel Core i7-8700K Processor, 64 GB DDR4 RAM, four GPUs Geforce GTX 1080-Ti and four GPUs NVIDIA Geforce RTX 2080-Ti. All the experiments were implemented in Python v3.6, and the CNN models were developed by using Keras framework v2.2.4 as high level API, and TensorFlow v1.12 as backend.

5. Results and discussion

This section analyzes the performance of the proposed framework in each step. Histograms and barcharts summarized the results obtained by the AL component of the framework.

5.1. Baseline methods

Firstly, the segmentation step was extensively evaluated by calculating the UN score of each image, and these results are available at the KDIS Research Group web page[3]. The applied segmentation method obtained the best average performance in the four datasets. Friedman's test rejected the null hypothesis in DERM-LIB, ISIC-2016 and ISIC-2017 datasets. The proposal was ranking first and afterwards, the Hommer's post-hoc test was conducted by considering our extension as the control method. The results showed the proposal significantly outperformed both U-Net and R2U-Net, except when comparing it to U-Net in ISIC-2017. In PH2 no significant differences were encountered between the three methods, the Friedman's test did not rejected the null hypothesis. However, the proposal was ranking first. The extension does not required prior training and can be used with only a CPU, avoiding a significant amount of computational power for training like those using GPU.

On the other hand, the selected CNN models were compared using and not the segmentation block in order to measure whether it is appropriate to use. DenseNet and MobileNet were the top average models, and achieved the best performance in 50% and 44% of the datasets, respectively. The 81% of the winners were the models that used the proposed segmentation method. NASNet reached 11% as maximum gain in MCC, as well as DenseNet, Xception, Inception, and MobileNet, with 7%, 6%, 6%, and 4%, respectively. In addition, we searched for significant differences between all the techniques grouped by CNN models. As a result, Friedman's test rejected the null hypothesis with a *p*-value of 3.606E-08, 8.355E-08, 7.356E-10, 1.161E-08 and 1.931E-06 in DenseNet, Inception, MobileNet, NASNet and Xception, respectively. The segmentation method was ranking first in all CNN models and afterwards, the Hommer's post-hoc test was conducted by considering STD as the control method. Our method significantly outperformed all techniques in Inception. In the rest of the models STD significantly outperformed baseline and data augmentation. Full results can be found in the available web page.

In summary, we corroborated that the segmentation method applied in this work was suitable to effectively improve the average prediction performance in all CNN models. In the following Sections we will compare the best

[3] Results can be found at http://uco.es/kdis/framework-cnn-melanoma

Table 3: Average MCC values obtained by using AL query strategies. The best MCC values by dataset were highlighted in bold typeface. "R", "L" and "M" mean Relevance Sampling, Least Confident and the proposal, respectively. The last rows show the Friedman's test rankings and the Hommel's p-values, where strike values represent the ones do not have significant differences compared to the best ranked.

Dataset	DenseNet			InceptionV3			MobileNet			NASNet			Xception		
	R	L	M	R	L	M	R	L	M	R	L	M	R	L	M
BCN20000	0.967	0.972	**0.991**	0.952	0.984	0.984	0.956	0.969	0.988	0.955	0.961	0.970	0.937	0.939	0.978
DERM-LIB	**1.000**	**1.000**	**1.000**	0.992	0.992	**1.000**	0.984	0.989	**1.000**	0.967	0.979	**1.000**	0.984	0.976	**1.000**
DERM7PT-C	0.858	0.891	**0.896**	0.812	0.820	0.844	0.838	0.884	0.889	0.722	0.780	0.886	0.716	0.777	0.818
DERM7PT-D	0.892	0.927	**0.936**	0.865	0.913	0.916	0.877	0.898	0.909	0.767	0.768	0.899	0.815	0.859	0.857
HAM10000	0.990	**1.000**	**1.000**	0.995	0.958	**1.000**	0.991	0.995	**1.000**	0.994	0.995	**1.000**	0.946	0.980	**1.000**
ISBI2016	0.948	**0.966**	0.927	0.954	0.955	0.955	0.947	0.953	0.950	0.922	0.937	0.940	0.922	0.918	0.920
ISBI2017	0.969	0.964	0.988	0.956	0.946	0.965	0.933	0.959	**1.000**	0.942	0.956	0.976	0.944	0.943	0.964
MED-NODE	0.856	0.879	0.887	0.861	0.924	**1.000**	0.857	0.858	0.887	0.771	0.740	0.887	0.799	0.837	0.849
MSK-1	0.949	**0.980**	0.977	0.934	0.953	0.977	0.964	0.973	0.951	0.867	0.883	0.932	0.881	0.888	0.878
MSK-2	0.966	0.962	0.982	0.938	0.940	0.964	0.934	0.989	**1.000**	0.919	0.931	0.870	0.964	0.943	0.964
MSK-3	0.980	**1.000**	**1.000**	0.959	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.842	0.854	**1.000**	**1.000**	**1.000**	**1.000**
MSK-4	0.957	0.963	0.971	0.936	0.941	0.919	0.944	**0.987**	0.961	0.853	0.896	0.941	0.914	0.936	0.913
PH2	**1.000**	**1.000**	**1.000**	0.973	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.955	0.891	**1.000**	0.984	0.984	**1.000**
SDC-198	**1.000**	**1.000**	**1.000**	0.984	0.984	**1.000**	**1.000**	**1.000**	0.979	0.905	0.914	**1.000**	0.968	0.969	**1.000**
UDA-1	0.889	0.906	0.913	0.877	0.862	0.864	0.875	0.853	**0.919**	0.685	0.758	0.760	0.782	0.820	0.778
UDA-2	0.747	0.778	0.800	0.676	0.701	0.707	0.739	0.784	0.707	0.683	0.581	0.707	0.424	0.505	**1.000**
Friedman	2.625	1.938	**1.438**	2.625	2.063	**1.313**	2.594	1.781	**1.625**	2.750	2.125	**1.125**	2.375	2.031	**1.594**
p-values	1.6E-3	~~1.6E-1~~	-	4.1E-4	3.4E-2	-	1.2E-2	~~6.6E-1~~	-	8.6E-6	4.7E-3	-	8.6E-6	4.7E-3	-

baseline and the best AL query strategy.

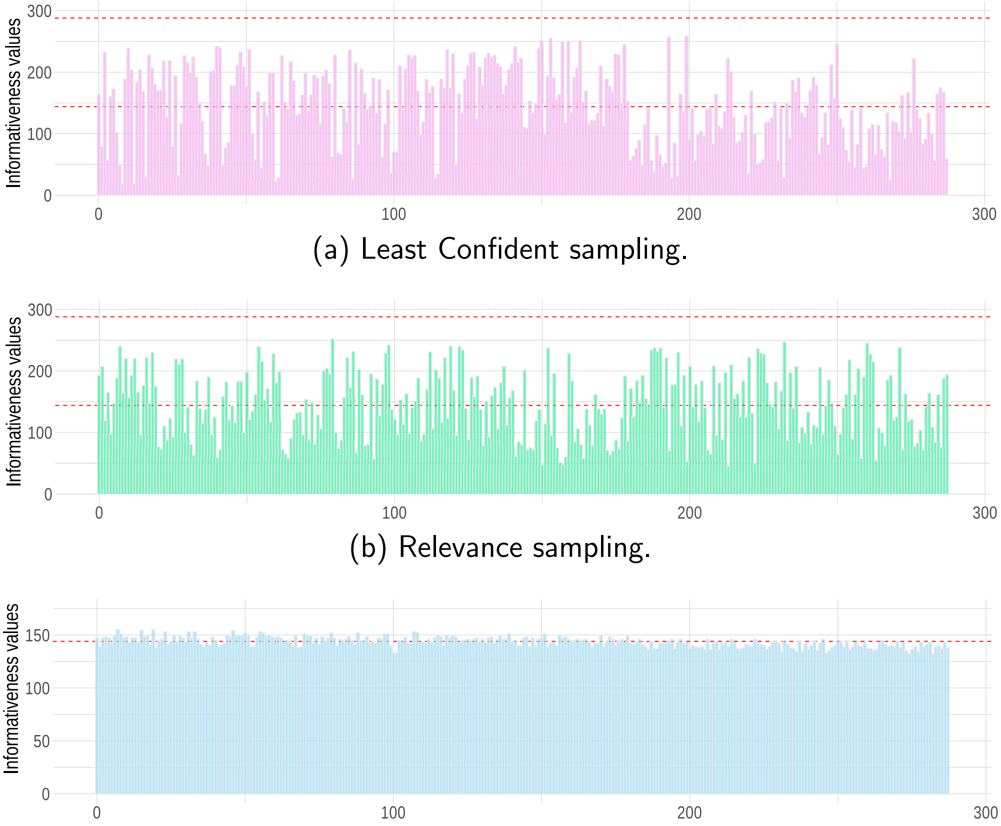
5.2. Comparison between AL query strategies

Table 3 shows the average MCC values obtained by all the AL query strategies considered in the comparative study. The results show that the proposed query strategy UMix achieved the maximum MCC value in nine datasets, and it did not deteriorate its performance on datasets having an imbalanced number of samples per class (e.g. MSK-3, HAM10000, and SDC-198). Statistical analysis were carried out independently for every CNN model. Friedman's test rejected the null hypothesis in all subsets of models and its rankings indicated the proposal obtained the first position always (see the penultimate row of the table). Afterwards, the Hommel's post-hoc test was conducted by considering the proposal as the control method, and the results showed it significantly outperformed Relevance Sampling in all CNN models and also Least Confident in InceptionV3, NASNet and Xception, so demonstrating that a superior predictive performance can be attained by applying new stable and more dynamic query strategies. The results revealed that AL is suitable to better train CNN models for melanoma diagnosis. However, it could be interesting to assess the nature behind how the query strategies drive the sampling of the data, which is analyzed in the next Section. These aspects help shed light on the AL component of the framework.

5.3. Analyzing the informativeness values

Figure 2 and Figure 3 show comprehensive views about the informativeness values from PH2 dataset, which guided the model through the sampling process. As can be seen in Figure 2, Least Confident and Relevance Sampling sampled images first more often than others along the dataset. The sampled images from both approaches followed different distributions, as expected due to their conflicting nature. However, bear in mind that although the both are opposites, the output rankings are not inversely equals due it depends in how each CNN model learns after each epoch. In fact, only the first ranking of both Least Confident and Relevance will be inversely equals. After the first epoch, the model will change its learning needs according with the given order. For example, Figure 3 shows a selection of 50 sorted samples. In all cases the first sample is the most informative, but the rest were selected at intervals of 5 samples in order to have an overall view about the sampling process. These values corroborate that all data are not seen more or less equally often in the same order (Figure 3a and Figure 3b). Also, it is shown the standard deviations from the sample indexes, which prove that after each epoch, φ changes its opinion about how informative an instance is, pointing to the fact that AL indeed helps φ during the training process. For example, the most informative sample in Least Confident is one of the less informative samples in Relevance Sampling. However, the most informative sample in Relevance is not the less informative in Least Confident. In fact, it is 40 positions ahead from the last ranking. Furthermore, Least Confident informativeness values fall faster and have less fluctuation compared to Relevance, e.g. the last sample in Figure 3a have an average informativeness value of 0.098 with a standard deviation of 27 compared to the last sample in Figure 3b with 0.185 and 71, which is almost the double and triple, respectively.

This analysis corroborated our initial hypothesis where φ learned about the complexities of individual samples and automatically changed its training sequence after each epoch. On top of that, a new query strategy was developed by combining both classical approaches and its informativeness values are also shown. Figure 2c and Figure 3c show that all instances are treated with similar importance at the end of the training process in our proposal. These three approaches corroborated that the most important point is to sample the instances according to some query strategy instead of random sampling. However, it could be interesting to assess how much time the framework requires to calculate the informativeness values after each epoch,

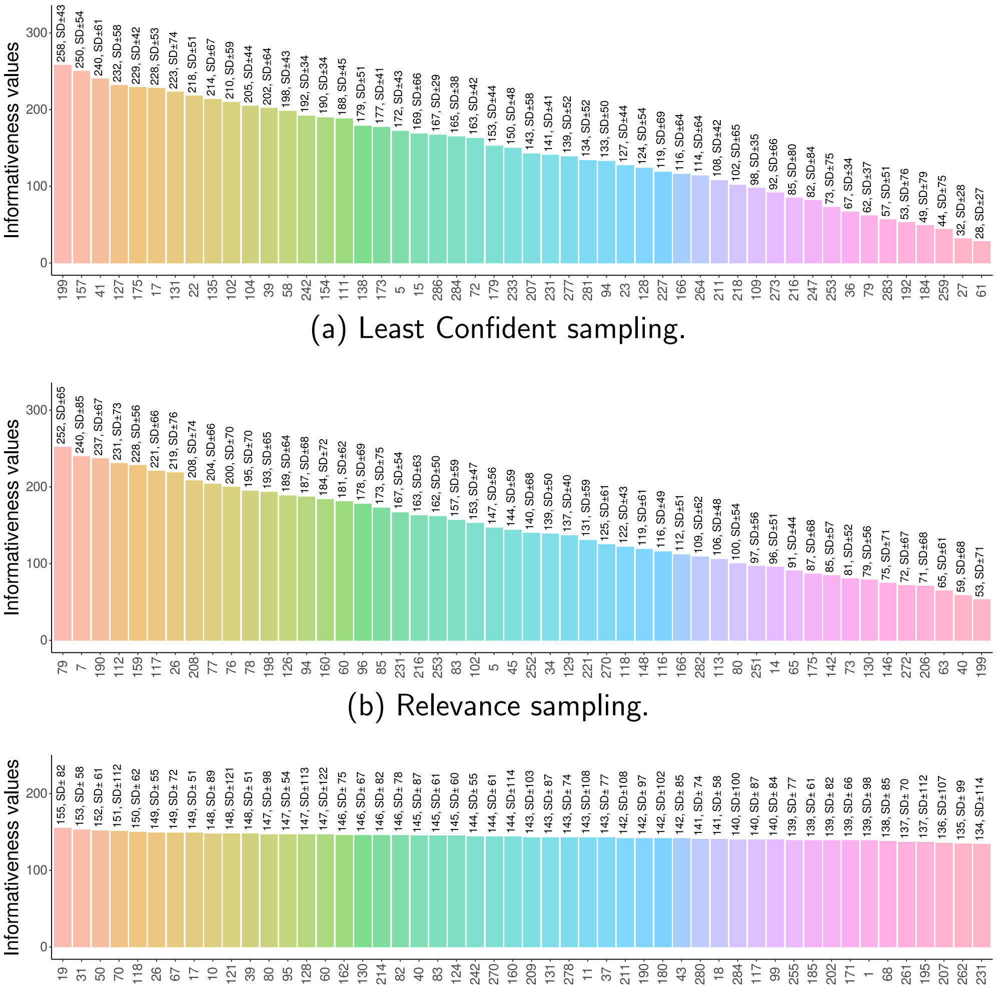


(a) Least Confident sampling.

(b) Relevance sampling.

(c) Proposed UMix sampling.

Figure 2: Example that shows the average informativeness value (Y axis) of every instance (X axis means the sample id) in PH2 dataset. Top dashed line means the maximum possible importance for this training set, while the medium dashed line represents equal average importance at the end of the training process.



(a) Least Confident sampling.

(b) Relevance sampling.

(c) Proposed UMix sampling.

Figure 3: Example showing details about the average informativeness values in a selection of 50 samples from PH2 dataset. It is shown the informativeness values rounded to the nearest integer (Y axis) and the standard deviation (SD); X axis means the sample id. The samples are ordered from left to right according to the informativeness values, the first samples are the most informative with 288 as the maximum possible value in this training set.

Table 4: Summary of the computational resources for training in PH2 dataset during 150 epochs. All columns indicate milliseconds, except GPU (GB). The best value for each column is highlighted in bold typeface, e.g. MobileNet is the one that requires less training time and finds faster the informativeness values.

CNN	Training	%	AL	%	Total	GPU
DenseNet201	9848550	92	845700	8	10694250	8.4
InceptionV3	4429500	92	401850	8	4831350	8.4
MobileNet	**1612050**	99	**21600**	1	**1633650**	8.1
NASNet	9918750	98	225150	2	10143900	**2.3**
Xception	3921150	87	593700	13	4514850	7.4

which is analyzed in the next Section.

5.4. Analyzing the computational cost

As for the computational cost of the proposed approach, Table 4 summarizes the training time of the compared models, as well as the proposed active learning approach. As mentioned before, the training time for segmenting is 0 ms, which is better compared to other previous proposals [32, 33, 34]. Nevertheless, the framework is flexible and allows to customize its internal steps, e.g. authors are allowed to change the segmentation method in step 1 to others such as U-Net and R2U-Net. Regarding the segmentation time, it took approximately 1713 ms to segment one image with only one CPU core. However, this process can be done only once at the beginning of the process. For example, the cost for segmenting all the 36,703 images from this work will be 17 hour/core. In this work we used two desktop Intel Core i7-8700K[4] (24 cores in total), being capable of segmenting all the samples in less than an hour. Regarding the proposed active learning framework, the computational time required by Q to calculate the informativeness values over all D depends on how complex the CNN model is. For example, by using MobileNet the proposed framework required only 1% of the total training time and achieved 8% better predictive performance, which is approximately 144 ms in each epoch in order to analyze the pool of samples. In addition, NASNet increased the performance by approximately 20% using non-dermoscopic images and the active learning process only required 2% of the total training time. Also, bearing in mind that AL is able to achieve its

[4] https://intel.ly/3hM0yaA

Table 5: Average MCC values obtained by comparing the best random baseline method ("STD") to the best AL query strategy ("UMix"). The best MCC values by dataset were highlighted in bold typeface; "F" represents the fold changes between the proposed query strategy and STD. The last row shows the Wilcoxon's test p-values.

Dataset	DenseNet			InceptionV3			MobileNet			NASNet			Xception		
	STD	UMix	F(%)	STD	UMix	F(%)	STD	UMix	F(%)	STD	UMix	F(%)	STD	UMix	F(%)
BCN20000	0.933	**0.991**	6	0.938	**0.984**	5	0.928	**0.988**	6	0.916	**0.970**	6	0.919	**0.978**	6
DERM-LIB	0.993	**1.000**	1	0.965	**1.000**	4	0.992	**1.000**	1	0.978	**1.000**	2	0.977	**1.000**	2
DERM7PT-C	0.813	**0.896**	10	0.771	**0.844**	9	0.812	**0.889**	9	0.728	**0.886**	22	0.741	**0.818**	10
DERM7PT-D	0.848	**0.936**	10	0.843	**0.916**	9	0.903	**0.909**	1	0.854	**0.899**	5	0.849	**0.857**	1
HAM10000	0.960	**1.000**	4	0.942	**1.000**	6	0.947	**1.000**	6	0.935	**1.000**	7	0.894	**1.000**	12
ISBI2016	0.878	**0.927**	6	0.825	**0.955**	16	0.854	**0.950**	11	0.829	**0.940**	13	0.843	**0.920**	9
ISBI2017	0.864	**0.988**	14	0.839	**0.965**	15	0.875	**1.000**	14	0.849	**0.976**	15	0.882	**0.964**	9
MED-NODE	0.699	**0.887**	27	0.766	**1.000**	31	0.768	**0.887**	15	0.665	**0.887**	33	0.759	**0.849**	12
MSK-1	0.940	**0.977**	4	0.873	**0.977**	12	0.899	**0.951**	6	0.852	**0.932**	9	0.860	**0.878**	2
MSK-2	0.889	**0.982**	10	0.817	**0.964**	18	0.870	**1.000**	15	0.867	**0.870**	0	0.832	**0.964**	16
MSK-3	0.959	**1.000**	4	0.969	**1.000**	3	0.966	**1.000**	4	0.808	**1.000**	24	0.938	**1.000**	7
MSK-4	0.868	**0.971**	12	0.847	**0.919**	9	0.850	**0.961**	13	0.823	**0.941**	14	0.829	**0.913**	10
PH2	0.967	**1.000**	3	0.964	**1.000**	4	0.987	**1.000**	1	0.937	**1.000**	7	0.971	**1.000**	3
SDC-198	0.956	**1.000**	5	0.929	**1.000**	8	**0.979**	**0.979**	0	0.922	**1.000**	8	0.941	**1.000**	6
UDA-1	0.808	**0.913**	13	0.762	**0.864**	13	0.813	**0.919**	13	0.716	**0.760**	6	0.734	**0.778**	6
UDA-2	0.513	**0.800**	56	0.410	**0.707**	72	0.580	**0.707**	22	0.441	**0.707**	60	0.355	**1.000**	182
p-values	2.189E-4		-	2.189E-4		-	2.412E-4		-	2.189E-4		-	2.189E-4		-

top predictive performance in fewer epochs, the true training time should be even lower. On the other hand, the most expensive for AL was DenseNet201, but only needing 8% of the total time in AL. In all cases the common training step consumed from 87% to 99% of the total time, which is expected. The proposal never consumed more than 13% of the total time. We believe our framework has an acceptable computational cost, mainly when it is combined with lightweight CNN models, such as MobileNet and NASNetMobile.

Next Section will conduct a comparison between the best baseline method and the best AL query strategy in order to conclude the advantages of this proposal. Bear in mind that the best baseline method was our proposed STD.

5.5. Comparing to state-of-the-art CNN models for melanoma diagnosis

Table 5 summarizes the results obtained after training the five CNN models with the baseline STD and UMix approaches. UMix obtained the highest MCC values in all datasets, except in SDC-198 and MED-NODE with MobileNet and Xception, respectively, where both methods tied. The results were very promising, because the proposed UMix outperformed STD in all the datasets. The smallest fold changes were observed on DERM-LIB and PH2 datasets with improvements of only 2% and 4% in average, respectively, and where the performance with both approaches is high. However, there were cases where the fold changes attained high rates, e.g. fold changes of

Table 6: Average epochs values obtained by comparing the best random baseline method to the best AL query strategy. The best epoch values by dataset were highlighted in bold typeface, smaller values mean a more desirable result. "STD" and "UMix" mean all the classical techniques and the proposed query strategy; "F" represents the fold changes, where negative values mean the percentage of epochs that UMix is better compared to STD, and positive values mean the opposite. The last row shows the Wilcoxon's test *p*-values.

Dataset	DenseNet			InceptionV3			MobileNet			NASNet			Xception		
	STD	UMix	F(%)	STD	UMix	F(%)	STD	UMix	F(%)	STD	UMix	F(%)	STD	UMix	F(%)
BCN20000	45	**14**	-69	44	**16**	-64	20	**16**	-20	39	**14**	-64	25	**9**	-64
DERM-LIB	27	**10**	-63	30	**10**	-67	20	**6**	-70	30	**10**	-67	27	**10**	-63
DERM7PT-C	**35**	36	3	38	**14**	-63	**37**	38	3	45	**24**	-47	45	**22**	-51
DERM7PT-D	**30**	40	33	42	**38**	-10	**32**	36	12	**45**	50	11	48	**19**	-60
HAM10000	60	**11**	-82	50	**12**	-76	20	**15**	-25	58	**10**	-83	17	**11**	-35
ISBI2016	**44**	45	2	50	**29**	-42	40	**20**	-50	75	**37**	-51	36	**23**	-36
ISBI2017	36	**13**	-64	32	**15**	-53	39	**19**	-51	75	**23**	-69	36	**15**	-58
MED-NODE	40	**8**	-80	42	**14**	-67	35	**13**	-63	40	**16**	-60	40	**15**	-62
MSK-1	69	**29**	-58	64	**23**	-64	45	**33**	-27	75	**26**	-65	43	**25**	-42
MSK-2	57	**40**	-30	62	**27**	-56	40	**19**	-52	79	**13**	-84	40	**15**	-62
MSK-3	32	**10**	-69	42	**16**	-62	30	**7**	-77	50	**11**	-78	40	**11**	-72
MSK-4	55	**32**	-42	55	**24**	-56	45	**11**	-76	65	**29**	-55	43	**16**	-63
PH2	25	**10**	-60	54	**11**	-80	30	**9**	-70	40	**10**	-75	39	**10**	-74
SDC-198	34	**14**	-59	59	**10**	-83	33	**10**	-70	62	**14**	-77	45	**14**	-69
UDA-1	45	**20**	-56	63	**30**	-52	45	**22**	-51	68	**13**	-81	52	**15**	-71
UDA-2	24	**18**	-25	41	**10**	-76	21	**11**	-48	16	**10**	-38	31	**10**	-68
p-values	8.046E-4		-	2.189E-4		-	4.262E-4		-	2.656E-4		-	2.189E-4		-

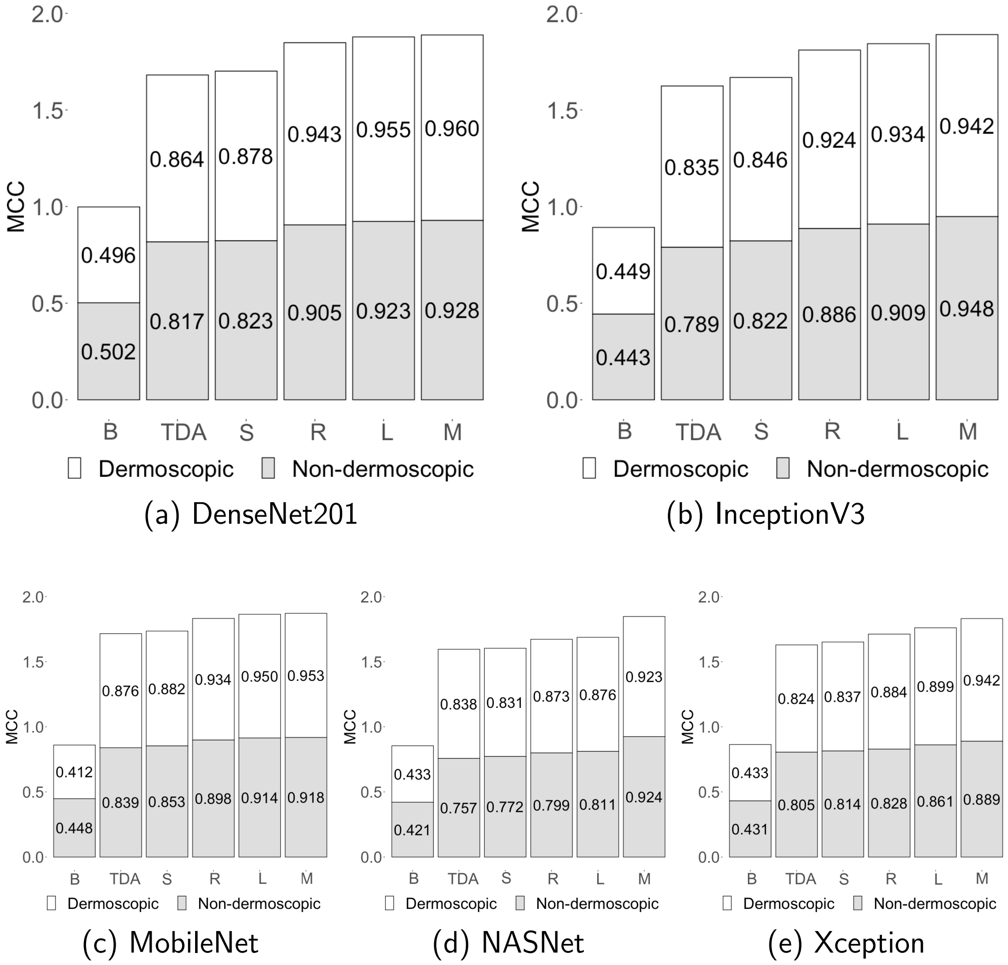
182% and 72% were obtained on the UDA-2 dataset with Xception and InceptionV3, respectively. In overall, the poorest predictive performance was observed on the UDA-2 dataset, in average the models achieved 78% MCC. UDA-2 has the lowest Silhouette value, indicating a high overlapping level between classes and increasing the difficulty. However, even so the proposed framework reached a performance 182% and 72% higher than when applying STD on Xception and InceptionV3, respectively. On the other hand, the best predictive performance was attained on the datasets DERM-LIB, HAM10000, PH2 and MSK-3. The first three datasets have the highest Silhouette values, meaning that they have a low overlapping level between images, which makes easier the task. Finally, in this comparison significant differences in performance were encountered, indicating the superiority of our proposal; the Wilcoxon's test rejected the null hypothesis in all CNN models and the *p*-values are shown in the last row. Results showed the proposed pipeline was able to selectively choose the order in which the model should visit the samples during training, and in this manner it was possible to obtain better performance in all CNN models. On the other hand, the number of epochs required to train the CNN models was analyzed; the aim was to find how quickly the CNN models achieved its top performance. Table 6 summarizes the results obtained after training the five CNN models. The results showed that all CNN models trained following the proposal framework UMix achieved the lowest number of epochs in the 92,5% of the time. DERM-LIB was the dataset where in average the models found faster their top performance needing only 9 epochs. In addition, DERM7PT-D and ISBI2016 were the most challenging datasets, in average the framework needed 37 and 31 epochs, respectively. In average, Xception and MobileNet are the fastest CNN models achieving its top performance, with an average of 16 and 19 epochs, respectively. Overall, DenseNet, InceptionV3, MobileNet, NASNet and Xception needed 45%, 61%, 46%, 61% and 59% less epochs to achieve their best predictive performance, respectively. Then, the Wilcoxon's test rejected the null hypothesis in all CNN models and the p-values are shown in the last row. The framework is not only capable of training accurately CNN models, but it also achieves the top performance in a faster way. Next Section will conduct a comparison between using or not dermoscopic images.

5.6. Dermoscopic versus non-dermoscopic images

Figure 4 shows the average performance attained by the models by grouping the datasets in dermoscopic and non-dermoscopic. This analysis shows differences between images taken with common digital cameras compared to those from dermatoscopes. Comparing the best baseline CNN models (S) versus the proposed UMix query strategy (M) by using non-dermoscopic images, it was obtained 20% and 15% of improvement on NASNet and InceptionV3, respectively. On the other hand, it was obtained 13%, 11% and 11% of improvement on Xception, InceptionV3 and NASNet by using dermoscopic images, respectively. Considering final results, DenseNet201 and InceptionV3 achieved the best performances on dermoscopic and non-dermoscopic images, respectively. The results showed that the proposed framework attained the best performance whatever the type of image, so denoting the effectiveness of the approach. In addition, regarding cost-effectiveness analysis, we believe the proposal is suitable. For example, training NASNet with the proposed framework increased the predictive performance in 20% and only consumed 2% of the total training time.

6. Conclusions

In this work, a framework for melanoma diagnosis has been proposed, allowing to build robust CNN models that best discriminate the images in their



(a) DenseNet201

(b) InceptionV3

(c) MobileNet

(d) NASNet

(e) Xception

Figure 4: Average MCC values on test sets by grouping the datasets in dermoscopic (top bars) and non-dermoscopic (bottom bars). “B”, “TDA”, “S”, “R”, “L” and “M” mean a baseline performance (i.e. none technique was applied), applying both transfer learning and data augmentation, combining all the above techniques with segmented images, Relevance Sampling, Least Confident and the proposal, respectively.

corresponding categories. In addition, we propose a new query strategy that achieves a more balanced training process. The proposed framework comprises three main phases, and we aim at showing the benefits of each block, being the proposed active learning query the main contribution. First, to segment data, we applied an extension of the Chan-Vese segmentation method. The segmentation masks indicated that the proposal achieved better performance compared to state-of-the-art biomedical segmentation methods. The selected CNN models achieved the best performance the 81% of the times by applying the segmentation block, which can be extended in other similar areas. According to the results, the method was significantly effective versus the rest of techniques when using InceptionV3. In addition, significant differences were also found in the rest of the models. Second, the data was increased to balance the number of images per category, reducing overfitting and obtaining transformation-invariant models. All models were benefit by using data augmentation, with a maximum gain of 683% in MobileNet, and an top average improvement of 122% in Xception. Also, we increased the performance applying transfer learning from pre-trained ImageNet. The applied weights have increased in a top of 88% the performance in MobileNet. Xception was the less benefited when using the above technique. Third, the CNN models were trained following an active learning mini-batch process, where the models were trained from images according to the order established by the query strategy. Not only all the models achieved significantly better performance compared to the their top baselines, but also they did it in less training epochs, as shown in Tables 5 and 6. In addition, the active component of the proposal could be seen in action when comparing the different query strategies. The CNN models and active learning were capable of changing the training sequence after each batch, which is the main contribution of the present work.

An extensive experimental study was conducted on sixteen image datasets, attaining significantly better results compared to the best baseline method. In addition, the best predictive performance was found in a significantly lower number of epochs. Finally, it is noteworthy that the proposed approach is not strictly restricted to melanoma diagnosis problem and could be applied in other areas. Future works will conduct more extensive experiments to validate the full potential of the framework, for example by considering a wide set of hyperparameters to be tuned as well as other more complex query strategies.

CRediT authorship contribution statement

Eduardo Pérez: Formal analysis, Investigation, Software, Validation, Original draft, Review and editing. **Sebastián Ventura**: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Supervision, Review and editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the Spanish Ministry of Science and Innovation, the University of Cordoba, and the European Regional Development Fund, projects UCO-FEDER 18 REF.1263116 MOD and PID2020-115832GB-I00. It was also funded by the i-PFIS contract no. IFI17/00015 granted by Health Institute Carlos III of Spain.

References

[1] American Cancer Society, Cancer Facts and Figures, 2021. URL: https://bit.ly/3gNDBVr, consulted on June 22, 2021.

[2] A. C. Geller et al., Screening, early detection, and trends for melanoma: Current status (2000-2006) and future directions, Journal of the American Academy of Dermatology 57 (2007) 555–572.

[3] N. R. Abbasi et al., Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria, Journal of the American Medical Association 292 (2004) 2771–2776.

[4] P. Carli et al., Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology, British Journal of Dermatology 148 (2003) 981–984.

[5] A. Esteva et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118.

[6] E. Pérez, O. Reyes, S. Ventura, Convolutional neural networks for the automatic diagnosis of melanoma: An extensive experimental study, Medical Image Analysis 67 (2021).

[7] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, volume 2, Harrahs and Harveys, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[8] R. Kaur, H. GholamHosseini, R. Sinha, M. Lindén, Automatic lesion segmentation using atrous convolutional deep neural networks in dermoscopic skin cancer images, BMC Medical Imaging 22 (2022) 1–13. URL: `https://doi.org/10.1186/s12880-022-00829-y`. doi:10.1186/s12880-022-00829-y.

[9] H. Basak, R. Kundu, R. Sarkar, MFSNet: A multi focus segmentation network for skin lesion segmentation, Pattern Recognition 128 (2022) 108673. URL: `https://doi.org/10.1016/j.patcog.2022.108673`. doi:10.1016/j.patcog.2022.108673. `arXiv:2203.14341`.

[10] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE transactions on pattern analysis and machine intelligence 43 (2019) 652–662.

[11] S. Lafraxo, M. E. Ansari, S. Charfi, MelaNet: an effective deep learning framework for melanoma detection using dermoscopic images, Multimedia Tools and Applications 81 (2022) 16021–16045. doi:10.1007/s11042-022-12521-y.

[12] U. Acharya, H. Fujita, O. Lih, M. Adam, J. Tan, C. Chua, Automated detection of coronary artery disease using different durations of ecg segments with convolutional neural network, Knowledge-Based Systems 132 (2017) 62–71.

[13] D. Gutman et al., Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC) (2016). URL: `http://arxiv.org/abs/1605.01397`.

[14] U. Asif, M. Bennamoun, F. Sohel, A Multi-Modal, Discriminative and Spatially Invariant CNN for RGB-D Object Labeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (2018) 2051–2065.

[15] Ericsson, On the pulse of the networked society, Technical Report, 2015. URL: `https://apo.org.au/node/59109`.

[16] K. Lenc, A. Vedaldi, Understanding Image Representations by Measuring Their Equivariance and Equivalence, International Journal of Computer Vision 127 (2019) 456–476.

[17] F. Perez et al., Data augmentation for skin lesion analysis, in: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, Springer, Granada, Spain, 2018, pp. 303–311.

[18] C. Baur, S. Albarqouni, N. Navab, MelanoGANs: high resolution skin lesion synthesis with GANs, arXiv preprint arXiv:1804.04338 (2018).

[19] R. Caruana, Multitask Learning, Machine Learning 28 (1997) 41–75.

[20] J. Kawahara et al., Seven-point checklist and skin lesion classification using multitask multimodal neural nets, IEEE Journal of Biomedical and Health Informatics 23 (2019) 538–546.

[21] D. Berthelot, T. Schumm, L. Metz, Began: Boundary equilibrium generative adversarial networks, 2017. `arXiv:1703.10717`.

[22] X. Li et al., Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation, IEEE Transactions on Neural Networks and Learning Systems 32 (2021) 523–534.

[23] O. Sener, S. Savarese, Active learning for convolutional neural networks: A core-set approach, in: 6th International Conference on Learning Representations, Vancouver, Canada, 2018.

[24] X. Shi et al., An active learning approach for reducing annotation cost in skin lesion analysis, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2019, pp. 628–636.

[25] Y. Leng, X. Xu, G. Qi, Combining active learning and semi-supervised learning to construct svm classifier, Knowledge-Based Systems 44 (2013) 121–131.

[26] O. Reyes et al., JCLAL: A Java framework for active learning, Journal of Machine Learning Research 17 (2016).

[27] T. Chan, L. Vese, An active contour model without edges, in: International Conference on Scale-Space Theories in Computer Vision, Springer, Corfu, Greece, 1999, pp. 141–151.

[28] N. Kowsalya et al., Skin-Melanoma Evaluation with Tsallis's Thresholding and Chan-Vese Approach, in: IEEE International Conference on System, Computation, Automation and Networking, ICSCA 2018, Pondicherry, India, 2018.

[29] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. `http://www.deeplearningbook.org`.

[30] X. Zhen et al., Handcrafted vs. learned representations for human action recognition, Image and Vision Computing 55 (2016) 39–41.

[31] L. Huang, Y.-G. Zhao, T.-J. Yang, Skin lesion segmentation using object scale-oriented fully convolutional neural networks, Signal, Image and Video Processing 13 (2019) 431–438.

[32] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, Munich, Germany, 2015, pp. 234–241.

[33] B. S. Lin et al., Skin lesion segmentation: U-Nets versus clustering, in: IEEE Symposium Series on Computational Intelligence, SSCI-2017, volume 2018-Janua, Hawaii, USA, 2018, pp. 1–7.

[34] M. Z. Alom et al., Recurrent residual U-Net for medical image segmentation, Journal of Medical Imaging 6 (2019).

[35] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: A survey, Knowledge-Based Systems 80 (2015) 14–23.

[36] M. A. Khan et al., Multi-model deep neural network based features extraction and optimal selection approach for skin lesion classification, in: Proc. of the International Conference on Computer and Information Sciences (ICCIS-2019), Karachi, Pakistan, 2019.

[37] I. Zliobaite et al., Active learning with drifting streaming data, IEEE Transactions on Neural Networks and Learning Systems 25 (2014) 27–39.

[38] M. Lin, K. Tang, X. Yao, Dynamic sampling approach to training neural networks for multiclass imbalance classification, IEEE Transactions on Neural Networks and Learning Systems 24 (2013) 647–660.

[39] D. Wilson, T. R. Martinez, The general inefficiency of batch training for gradient descent learning, Neural Networks 16 (2003) 1429–1451.

[40] D. Mumford, J. Shah, Optimal approximations by piecewise smooth functions and associated variational problems, Communications on Pure and Applied Mathematics 42 (1989) 577–685.

[41] K. Suzuki et al., Machine Learning in Medical Imaging: Second International Workshop, MLMI 2011, Held in Conjunction with MICCAI 2011, volume 7009, Springer, Toronto, Canada, 2011.

[42] M. Ahmad et al., Spatial prior fuzziness pool-based interactive classification of hyperspectral images, Remote Sensing 11 (2019).

[43] D. D. Lewis, W. A. Gale, A Sequential Algorithm for Training Text Classifiers, in: Proc. of the ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 1994, pp. 3–12.

[44] D. D. Lewis, J. Catlett, Heterogeneous uncertainty sampling for supervised learning, in: Proc. of the International Conference on Machine Learning (ICML), Morgan Kaufmann, San Francisco, USA, 1994, pp. 148–156.

[45] H. S. Seung, M. Opper, H. Sompolinsky, Query by Committee, in: Proc. of the ACM Workshop on Computational Learning Theory, Pennsylvania, USA, 1992, pp. 287–294.

[46] B. Settles, M. Craven, An Analysis of Active Learning Strategies for Sequence Labeling Tasks, in: Proc. of the Conference on Empirical Methods in Natural Language Processing, October, Honolulu, Hawaii, 2008, pp. 1070–1079.

[47] E. Simo-Serra et al., Discriminative learning of deep convolutional feature point descriptors, in: Proc. of the IEEE International Conference on Computer Vision, 2015, pp. 118–126.

[48] Y. Bengio et al., Curriculum learning, in: Proc. of the 26th International Conference On Machine Learning, ICML 2009, 2009, pp. 41–48.

[49] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.

[50] T. Mendonca et al., Ph2 - a dermoscopic image database for research and benchmarking, in: Proc. of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Osaka, Japan, 2013, pp. 5437–5440.

[51] L. Ballerini et al., A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions, volume 6, 2013.

[52] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proc. of the 13th International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 2010, pp. 249–256.

[53] I. Goodfellow et al., Deep learning, volume 1, MIT press Cambridge, 2016.

[54] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (2012) 281–305.

[55] S. Patidar, R. Pachori, U. Rajendra Acharya, Automated diagnosis of coronary artery disease using tunable-q wavelet transform applied on heart rate signals, Knowledge-Based Systems 82 (2015) 1–10.

[56] S. Boughorbel et al., Optimal classifier for imbalanced data using matthews correlation coefficient metric, PloS one 12 (2017) e0177678.

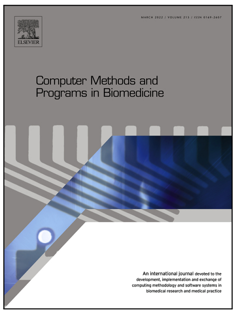
[57] T. Gross, M. Bessani, W. Darwin Junior, R. Araújo, F. Vale, C. Maciel, An analytical threshold for combining bayesian networks, Knowledge-Based Systems 175 (2019) 36–49.

[58] E. Pérez, S. Ventura, Melanoma recognition by fusing convolutional blocks and dynamic routing between capsules, Cancers 13 (2021). doi:10.3390/cancers13194974.

[59] G. Hommel, A stagewise rejective multiple test procedure based on a modified bonferroni test, Biometrika 75 (1988) 383–386.

[60] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945) 80–83.

6.5. Generative Adversarial Networks


MARCH 2022 / VOLUME 215 / ISSN 0169-2607
ELSEVIER
Computer Methods and
Programs in Biomedicine
An international journal devoted to the
development, implementation and exchange of
computing methodology and software systems in
biomedical research and medical practice

Title	Progressive growing of Generative Adversarial Networks for improving data augmentation and skin cancer diagnosis
Authors	E. Pérez, and S. Ventura
Journal	Artificial Intelligence in Medicine
Status	Submitted
Year	2022
Editorial	Elsevier

IF (JCR 2021)	7.011
Categories	Computer Science - Artificial Intelligence, Medical Information
Positions	32/145 (Q1), 8/31 (Q2)

Progressive growing of Generative Adversarial Networks for improving data augmentation and skin cancer diagnosis

Eduardo Pérez[a,b], Sebastián Ventura[a,b,c,*]

[a] *Department of Computer Science and Numerical Analysis, University of Córdoba, Córdoba, Spain*
[b] *Maimonides Biomedical Research Institute of Córdoba, Spain*
[c] *Department of Information Systems, King Abdulaziz University, Saudi Arabia*

Abstract

Early melanoma diagnosis is the most important factor in the treatment of skin cancer and can effectively reduce mortality rates. Recently, Generative Adversarial Networks have been used to augment data, prevent overfitting and improve the diagnostic capacity of models. However, its application remains a challenging task due to the high levels of inter and intra-class variance seen in skin images, limited amounts of data, and model instability. We present a more robust Progressive Growing of Adversarial Networks based on residual learning, which is highly recommended to ease the training of deep networks. The stability of the training process was increased by receiving additional inputs from preceding blocks. The architecture is able to produce plausible photorealistic synthetic 512 × 512 skin images, even with small dermoscopic and non-dermoscopic skin image datasets as problem domains. In this manner, we tackle the lack of data and the imbalance problems. Additionally, the proposed approach leverages a skin lesion boundary segmentation algorithm and transfer learning to enhance the diagnosis of melanoma. Inception score and Matthews Correlation Coefficient were used to measure the performance of the models. The architecture was evaluated qualitatively and quantitatively through the use of an extensive experimental study on sixteen datasets, illustrating its effectiveness in the diagnosis of melanoma. Finally, four state-of-the-art data augmentation techniques applied in five convolutional neural network models were significantly outperformed. The results indicated that a bigger number of trainable parameters will not necessarily obtain a better performance in melanoma diagnosis.

Keywords:
Melanoma diagnosis, Generative Adversarial Networks, Residual connections, Transfer learning

1. Introduction

Melanoma is one of the most lethal types of skin cancer and begins in cells known as melanocytes. According to the American Cancer Society in its latest publication of "Cancer Facts & Figures 2021", just in the United States, more than 100,000 new cases of melanoma and 7,000 deaths were expected in 2021 [1]. However, the early diagnosis can increase the chance of survival, achieving a 98% five-year survival rate. The first step in the diagnosis of a skin lesion by a dermatologist is an initial clinical examination. If any doubt remains, a dermoscopic analysis, biopsy and histopathological examination are performed [2]. With regard to clinical diagnosis, dermatologists have an accuracy rate of between 65%-80% and up to 75%-84% through the use of dermoscopic images [3, 4]. Despite the expertise of dermatologists, early diagnosis of melanoma remains a daunting task as it presents in many different shapes, sizes and colors - even between samples in the same category.

Nowadays, advanced computational techniques are used in the diagnosis of melanoma in order to make the diagnosis easier and aid dermatologists in their decision making. For example, those based on descriptors [5] and Convolutional Neural Networks (CNNs) [6]. Descriptor-based methods require the previous extraction of handcrafted features, which involves the expertise of dermatologists. However, this task is time-consuming and prone to errors [7]. In order to solve such limitations, CNN models have been applied to learn high-level features from raw images without the involvement of experts [2]. Several authors have cor-

[*] Corresponding author
Email address: `sventura@uco.es` (Sebastián Ventura)

roborated that CNN models can overcome handcrafted feature-based methods [8] and can even rival the prediction accuracy of dermatologists [9].

CNN models have proven effective in solving complex problems [10]. However, these models present several disadvantages, specifically regarding their application to skin lesion diagnosis. They are prone to overfitting on datasets with a small number of training examples per category and, as a result, attaining a poor generalization capacity. Also, CNNs require large datasets in order to learn accurately, which is a major issue in public melanoma datasets. On top of that, most datasets are unbalanced, and the minority category is often melanoma. On the other hand, CNN models are sensitive to some characteristics in data, such as large inter-class similarities and intra-class variances, variations in viewpoints, changes in lighting conditions, occlusions and background clutter [11].

CNN models seem to work better with high-quality and standardized conditions, such as dermoscopic images. However, keeping in mind the growing tendency to collect images taken with common digital cameras, we made the effort to include as many non-dermoscopic datasets as possible [12]. In this way, it is possible to reduce the number of invasive treatments and required economic resources in addition to boosting the development of modern and inexpensive tools. Finally, CNN models are approximately invariant to small translations to the input but are not rotation, color or lighting-invariant [13]; invariance to a transformation is an important concept in the realm of image recognition. If you take the input and transform it, the representation you end up with is the same as the representation of the original.

There are several techniques for overcoming the problem of invariance and limited training datasets. The most simple include basic data augmentation [14], transfer learning [2] and ensemble learning [15]. In the past few years, other techniques such as advanced data augmentation through the use of multi-task learning (MTL) [16] and Generative Adversarial Networks (GANs) [17], have emerged. Also, new architectures have been proposed based on representing and preserving properties of a object such as position, size, texture, and hierarchical spatial relationships [18]. So far, basic techniques have been widely applied in the diagnosis of melanoma, significantly improving the performance of CNN models. However, advanced techniques like GANs have been used with discretion in the diagnosis of melanoma, which is further explained in Section 2.1. Consequently, in this work, a new GAN-based approach for diagnosing melanoma from images and a deep analysis of its core components are carried out.

Firstly, a more robust architecture inspired by GANs is proposed. Generally speaking, GANs consists of two models that are trained together in an adversarial zero-sum game: a generative model that captures the data distribution and a discriminative model that predicts whether a sample is fake or not. However, GANs are limited to small image sizes due to model stability. We propose a customized Progressive Growing GAN architecture (PGGAN), which is a stable approach for training GANs models to generate large high-quality images. This involves incrementally increasing the size of the model during training and requires a highly complex training process. In addition, bearing in mind the restrictions in our problem, where only few images are commonly available and high-resolution images are needed, two residual connections were used to progressively train the models, instead of one. This approach is inspired by Residual Learning Framework (ResNet) [19] and Dense Convolutional Network (DenseNet) [20], where each building block receives additional inputs from preceding blocks. By doing this, the stability of the training process was further increased, particularly in small datasets. As such, the generative model is capable of generating photorealistic synthetic 512×512 skin images which are almost indistinguishable from the real ones. As a result, it is possible to accurately train deep CNN models to overcome the lack of data present in skin datasets. Additionally, data augmentation helps to improve the generalization capacity of CNN models and can be used as a regularization method in order to prevent overfitting [21]. It is interesting to compare basic and advanced data augmentation techniques in a large number of datasets. Secondly, a lesion segmentation method based on Chan-Vese segmentation algorithm is applied [15]. It is well understood that a preprocessing phase can improve the quality of any biomedical data analysis [22]. The algorithm is capable of effectively obtaining reliable segmentation masks without prior knowledge. Thirdly, the proposed double residual architecture was assessed qualitatively and quantitatively by using state-of-the-art metrics. In this manner, we double-checked for realism and high-quality images. Then, an extensive experimental study was conducted on sixteen skin image datasets in order to evaluate the augmented images. Five CNN models were trained with several state-of-the-art data augmentation techniques, and that performances were compared when training with the augmented images. The results showed that the proposed approach attained suitable results and significantly outperformed the rest of techniques.

The rest of this work is organized as follows: Section 2.1 briefly presents the state-of-the-art in solving melanoma diagnosis problem mainly by using CNN models; Section 2.2 describes the proposal architecture to augment images and how the CNN models were trained; the analysis and discussion of the results are portrayed in Section 3; finally, the concluding remarks are presented in Section 4.

2. Material and methods

This section firstly describes the related works regarding the automatic cancer diagnosis from image data, and then it presents the proposed architecture for data augmenting images.

2.1. Related works

Since 2016, *The International Skin Imaging Collaboration*[1] (ISIC) project organizes annually a challenge in which more than 180 teams have already participated. Most submissions are based on CNN models, and it is very common that authors applied additional techniques to even improve the performance, such as transfer learning [2], lesion segmentation [23], data augmentation [14] and in the last years advanced data augmentation by using GANs [24].

Transfer learning is an effective method that tries to transfer and reuse the knowledge that was extracted from a source task, on a target task. The above helps to alleviate the fact that it is required an enormous collection of data in order to build accurate CNN models. For example, Esteva et al. [2] used Google's InceptionV3 architecture pretrained on ImageNet, they remove the final classification layer and then they re-trained with 129,450 skin lesion images. The CNN achieved a great performance as was mentioned before. In addition, Liang and Zheng [25] applied a custom transfer learning method for pediatric pneumonia diagnosis, which involved 112,120 chest X-ray images labeled with 14 different chest diseases. The authors explored the problem of low image resolution, partial occlusion, and the importance of transfer learning. The results showed that the proposal achieved the top diagnostic performance in the classification task of children pneumonia.

On the other hand, skin lesion segmentation is able to isolate the lesion and it plays an important role improving the performance of CNN models. This is a complex task and it is very important because some areas not related with the lesion can lead CNN models to misclassify samples. From ISIC-2016 to ISIC-2018 there was a special task related to lesion segmentation. In ISIC-2016, considering the top three average results, there was a slight improvement when using segmentation.

Data augmentation is employed to obtain new data through transformations of existing images [6] or generating new ones based in such images [24]. The above helps reducing overfitting and obtaining transformation-invariant models [13]. In the case of melanoma diagnosis, most of the datasets available lack of balance between the categories, so data augmentation is commonly used to tackle imbalance [14]. For example, Esteva et al [2] showed the suitability of CNN models as a powerful tool for melanoma diagnosis. The authors augmented the images by a factor of 720 using basic random transformations, such as rotation, flip and crop. Also, they compared the performance of a CNN model versus 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. The results showed that the CNN model achieved a performance on par with experts in both tasks. In addition, Lenc and Vedald [13] trained CNN models by applying random data augmentation, where the benefits were more noticeable in deeper models.

Recently, advanced data augmentation techniques have been proposed, such as GANs, Random Erasing Data Augmentation (RE) [26] and Unsupervised Data Augmentation for Consistency Training (UDA) [27]. GANs is a relative new technique to augment data. GANs represents a way of training two models: the generator model that it is trained to generate new examples, and the discriminator model that tries to classify examples as real or fake. It is commonly used to generate new images that plausibly could have been drawn from the original dataset. Baur et al [17] applied GANs to generate realistically looking high resolution images of skin lesions. The authors proposed a new GANs model based on two classical GANs architectures, namely Deep Convolutional GANs [28] and Laplacian Pyramid of Adversarial Networks [29], and they showed that this type of method are able to mimic the data distribution with diverse and realistic samples, even when the training dataset is very small. In addition, Qin et al. [30] proposed a custom style-based GAN architecture in order to generate 256×256 images, where the same weights were injected directly in the growing layers. The dataset of the International Skin Imag-

[1] https://www.isic-archive.com

ing Collaboration Challenge (2018) was used as source data. The authors concluded that the synthetic images helped the diagnostic model to achieve a better classification performance. Nevertheless, we believe that there is still room for further study of GAN architectures in the diagnosis of melanoma. For example, although there are some studies, they are limited to corroborating their proposal on only one dataset, which can be a constraint when applying a proposal on another dataset with different characteristics. In addition, to the best of our knowledge, there is not evidence of the analysis of GAN in non-dermoscopic images.

On the other hand, Random Erasing generates images with various levels of occlusion by randomly erasing pixels in an image, which reduces the risk of overfitting. This technique do not have parameter and is easy to implement. The method obtained good performance in several datasets, such as CIFAR10, CIFAR100, and Fashion-MNIST. The method even achieved a reasonable improvement on object detection and person reidentification. In Unsupervised Augmentation, the authors investigated the noise injection during training, and concluded that RandAugment [31] and back-translation [32] methods achieved competitive performance compared to other techniques.

Despite the advantages that some techniques offer for a better training of CNN models, such as ensemble and multi-task learning methods, it should be stressed that important limitations arise when they are applied. For example, the training of ensemble commonly requires the assessment of a high number of possibilities that hamper the process, and the way to combine the learned representations and partial predictions yielded by each member of an ensemble. On the other hand, the main limitation when applying MTL is the lack of public datasets that contain heterogeneous information, e.g image and clinical data of each patient.

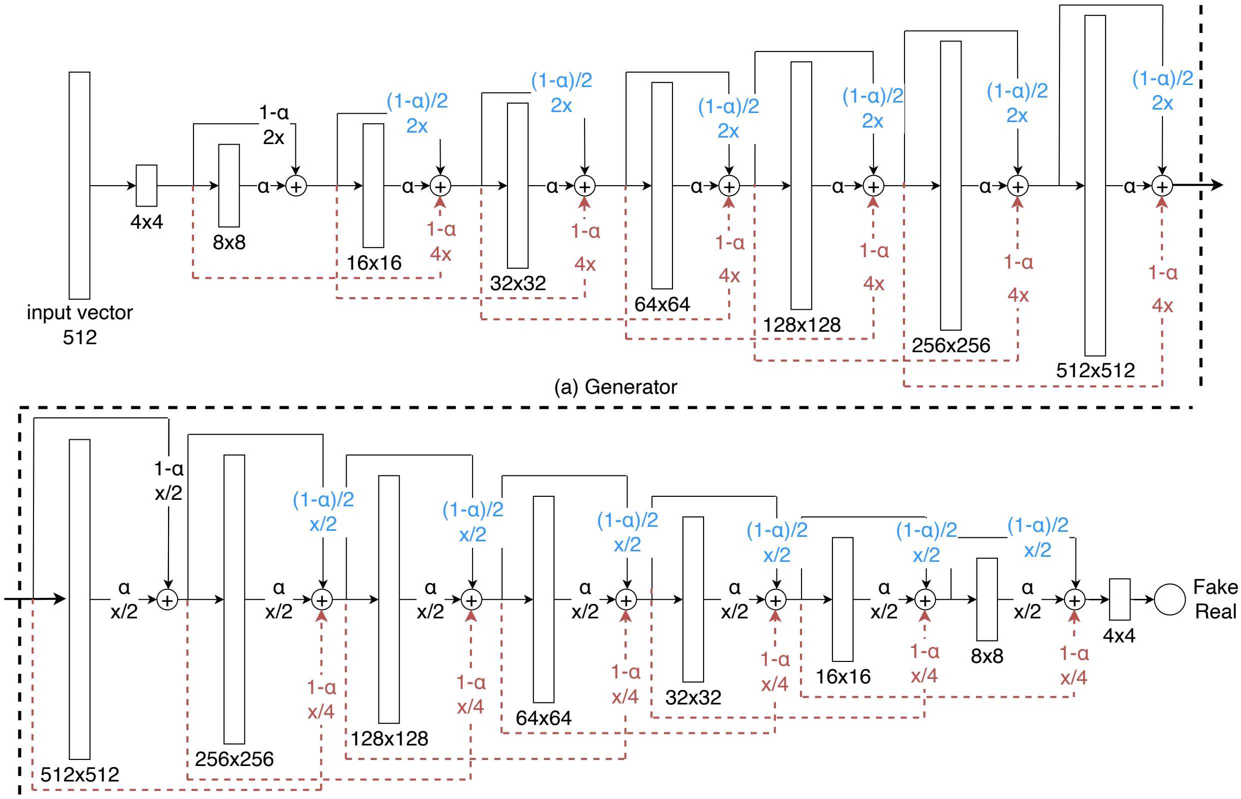
To sum up, in this work our main aim is to explore how to generate plausible high resolution skin images by designing a customized PGGAN architecture. The proposal applies two residual connections in order to progressively train the models, which is a valid approach to increase the stability of the training process. Also, we evaluate the performance of CNN models with augmented data by applying basic techniques versus advanced ones in an extensive experimental study conducted on sixteen image datasets. So far, there is no evidence of previous research addressing the increase of residual connections and conducting such an in-depth evaluation. In addition, transfer learning and segmentation to improve the performance of CNN models are applied. In the next section, the proposed architecture is described.

2.2. *Double residual Progressive Growing of Generative Adversarial Networks*

GANs was first described in 2014 by Ian Goodfellow [33]. The proposal was composed by two sub-models, a generator used to generate new believable examples from a domain dataset and a discriminator used to classify each one as real or fake. The generator tries to fool the discriminator by generating plausible images. Then, the parameters of the generator are updated by using the feedback from the discriminator. Commonly, when the discriminator is fooled about half the time, means the generator is ready. The architecture was unstable and hard to train at the beginning. However, Radford et al. [28] proposed a stable approach called Deep Convolutional Generative Adversarial Networks (DCGAN), and nowadays most GANs architectures are based on it. Consequently, GANs has started to be widely applied in tasks related with images, especially to augment data when training datasets are limited.

Traditional GANs models work well in low resolution datasets, commonly less than 100-pixel square images, which can be a problem if the problem domain comprises medium resolution images, such as dermoscopic or non-dermoscopic skin images. Generating high-quality images is one of the challenge for GANs, due the generator must learn how to output a large structure and small details. High resolution images make easy to detect issues in the fake images for the discriminator, so the training process can fail. Also, large images require more GPU memory, which is a disadvantage if you are using enthusiastic GPU cards. Therefore, the batch size to update the model weights each training iteration must be reduced to fit into memory. The above introduces instability into the training process. Karras et al. [34] proposed a solution for high-resolution generative models, which consisted in progressively increase the number of layers during the training process. This approach is called Progressive Growing GAN. The incremental addition of layers allows the models to discover large-scale structure of the image distribution and then it focuses on to increasingly finer scale detail, instead of learning all scales at the same time. In this work, we generate realistically looking high resolution skin images from small and medium dataset sizes.

Figure 1 shows the steps during the transition between each block of layers in the proposed architecture for melanoma diagnosis. Firstly, the generator requires points in a specific latent space to generate new output images. The generator will give meaning to the latent



(a) Generator

(b) Discriminator

Figure 1: Transition between each layer of the proposed generator and discriminator models. The input vector is formed by sampling from a standard Gaussian distribution; 2× and 4× mean doubling and quadrupling the image resolution in the generator, and $x/2$ and $x/4$ represent the opposite, respectively. The colors represent the introduced changes. Red color represents the new injected outputs and weights, and blue color represents the modified weights.

points and at the end of training, the latent space represents a compressed representation of the output space. The more recent best practice is to sample from a standard Gaussian distribution, meaning that the shape of the latent space is a hypersphere, with a mean of zero and a standard deviation of one [35]. There is not a specific dimension for the latent space, but it is recommended to use values between 100 and 512.

Secondly, the process of growing GANs architectures requires adding layers to the generator and discriminator during the training process, specifically blocks of layers. The blocks are phased in the addition of the blocks of layers rather than adding them directly. In this work, the output of each layer is modified by the output of the previous ones, by using a double residual connection. This approach is inspired by ResNet and DenseNet, where each building block receives additional inputs from preceding blocks (Figure 2). Residual learning is highly recommended for easing the training of deep networks, which is considered in this work [19]. By doing this, the stability of the training process was further increased.

In addition, bearing in mind that most of the existing melanoma datasets only encompass a few hundred of images, a second residual connection was used. The main output is weighted by α, the intermediate one by $\frac{1-\alpha}{2}$ and the most scaled-up output by $1-\alpha$; α is small initially, giving first the biggest weight to the largest scaled-up version of the image, although slowly transitions to giving more weight and then all weights to the new main output layers over training iterations.

The generator starts with a very low resolution image, just about 4×4 and ends up with 512×512, which is the final image resolution (see Figure 1a). We use leaky ReLU in all layers of generator and discriminator, except for the last layer that uses linear activation. Nearest neighbor filtering was used for increasing the image resolution. The resolution is chosen taking into account that some well-known models, such as InceptionV3, require image quality higher than 256×256. The purpose

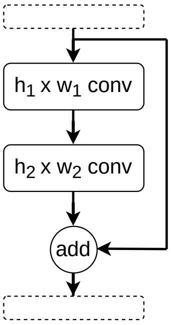


Figure 2: Example of a residual block.

is not to lose image quality in any case. On the other hand, the discriminator takes as input an image from the generator and outputs *fake* or *real*. The discriminator executes the opposite process of the generator; now the goal is to downsample the image progressively until a resolution of 4 × 4 is attained (see Figure 1b). Average pooling was applied when downsampling the image. The downsampled versions of the input are progressively combined in a weighted manner, in a similar way as the generator. In this work, eight blocks of layers for both generator and discriminator models were used.

The classical GANs architecture is trained by using a minimax GANs loss; minimax loss means the minimization of the generator and the maximization of the discriminator's loss. The loss produced by both models can be calculated as follows,

$$\mathcal{L}_D^{GAN} = -\mathbb{E}_{x\sim\mathbb{P}_d}[log(D(x))] - \mathbb{E}_{\hat{x}\sim\mathbb{P}_g}[log(1 - D(\hat{x}))], \quad (1)$$

$$\mathcal{L}_G^{GAN} = \mathbb{E}_{\hat{x}\sim\mathbb{P}_g}[log(1 - D(\hat{x})], \quad (2)$$

where $\mathbb{P}_d$ is the data distribution and $\mathbb{P}_g$ is the model distribution; x is a real image and $\hat{x} = G(z)$ where z is a standard Gaussian distribution and $\hat{x}$ is a synthetic image; D and G mean the discriminator and generator, respectively. In some scenarios it was found that if the generator cannot learn as quickly as the discriminator, the generator's loss saturates and the discriminator wins [33].

Consequently, nowadays there are other loss formulations designed to overcome the saturation problem. The least squares (LSGAN) [36] and Wasserstein loss (WGAN) [37] functions are commonly used in modern GANs architectures. Gulrajani et al. [38] demonstrated that both loss functions obtained good performance, however, Wasserstein loss was better compared to least squares for PGGAN architectures. These losses can be calculated as

$$\mathcal{L}_D^{LSGAN} = -\mathbb{E}_{x\sim\mathbb{P}_d}[(D(x) - 1)^2] + \mathbb{E}_{\hat{x}\sim\mathbb{P}_g}[D(\hat{x})^2], \quad (3)$$

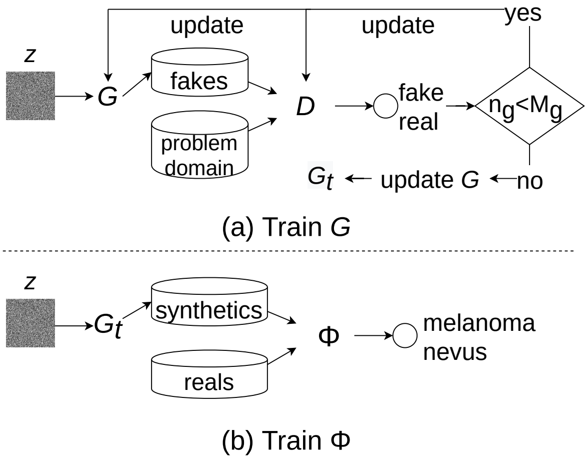


Figure 3: Proposed pipeline for melanoma diagnosis; G, D and G_t mean the generator, the discriminator and the trained generator, respectively; "fakes" and "reals" are the images created by the generator and the original ones, respectively; Φ is a CNN model trained for identifying if an image is melanoma or nevus; n_g and M_g are the number of epochs trained so far and the maximum number of epochs, respectively.

$$\mathcal{L}_G^{LSGAN} = -\mathbb{E}_{\hat{x}\sim\mathbb{P}_g}[(D(\hat{x} - 1))^2], \quad (4)$$

$$\mathcal{L}_D^{WGAN} = -\mathbb{E}_{x\sim\mathbb{P}_d}[D(x)] + \mathbb{E}_{\hat{x}\sim\mathbb{P}_g}[D(\hat{x})], \quad (5)$$

$$\mathcal{L}_G^{WGAN} = -\mathbb{E}_{\hat{x}\sim\mathbb{P}_g}[D(\hat{x})]. \quad (6)$$

Wasserstein loss provides a useful gradient almost everywhere, allowing for the continued training of the models. Also, a lower Wasserstein loss correlates with a better generator image quality, so we are seeking for a minimization in the generator loss. The above loss function was the first one showing this property. To sum up, Wasserstein loss shows the greatest advantages and it is applied in the proposed PGGAN architecture for melanoma diagnosis. Next, it is shown how to combine the PGGAN components mentioned so far.

Figure 3a shows how the proposed architecture is trained. The algorithm involves training both generator and discriminator in parallel. In addition, the training process requires a problem domain image dataset, from which synthetic images will be generated. Commonly, skin image datasets are unbalanced at the expense of the melanoma category. As a result, melanoma images were taken as problem domain in almost all datasets when augmenting the images to feed the CNN models. However, both normal and melanoma images were assessed when evaluating the generative architecture in Section 3.5.1. The aim is to increase the minority category in order to balance the training set while keeping low the number of images and the required computational resources. First, the latent space z is generated following a standard Gaussian distribution. Then, all pixel values from images are transformed in the range $[-1, 1]$ before passing to the models [28]. Second, $G(z)$ outputs fake images. Third, D is trained with fake and real images, which is recommended to do it with separated batches; first by using real images and then by using only the fake ones [35]. Four, the weights are updated. The last three steps are repeated until a stop condition is reached. In our case, the number of epochs and the Inception Score were selected as stop criterion. Finally, the generator associated to the problem domain is obtained. Generally speaking, the generator is capable of generating suitable images between epochs 100 and 300, and also between 300 and 450 as well. In this work, we noticed that training between 190 and 260 epochs was enough to obtain an acceptable performance in all skin image datasets. Next, the above generator is used to create synthetic images, which are used to train all CNN models.

Figure 3b shows how a deep CNN model is trained in order to predict whether an image is melanoma or not. Let us say T is a dataset of synthetic (s) and real (r) images, where x_i represents the ith image and y_i its label. In this work two classes were considered, melanoma ($y = 1$) and nevus ($y = 0$). Let be Φ a model that follows a CNN architecture, which learns the representations from the feature space and yields a prediction. Φ extracts representations from the dataset and finally, Φ predicts the label of the sample i ($\hat{y}_i$). Once the prediction for a given training image is computed, the loss obtained by applying Φ on the ith training image ($\mathcal{L}(i)$) is computed by means of a binary cross entropy.

Finally, in this work the CNN models were trained using mini-batch gradient descent. This method splits the training dataset into small batches that are used to calculate the model error. The above method have several advantages, such as the model update frequency is higher than batch gradient descent, which allows for a more robust convergence, avoiding local minima. In addition, batch-based updates provide a computationally more efficient process than stochastic gradient descent; and the split in small batches allows the efficiency of not having all training data in memory. All images in the batch are processed in parallel using GPU memory, increasing significantly the training speed; also, small batches can serve as regularizing effect.

3. Experimental study

This section describes the experimental study carried out in this work. First, the datasets and the experimental protocol are portrayed, and then, the experimental results and a discussion of them are presented.

Table 1: Summary of the benchmark datasets.

Dataset	Img	ImbR	IntraC	InterC	DistR	Silho
BCN20000	17,393	2.848	9,014	10,107	0.892	0.153
DERM-LIB	407	4.355	7,171	9,163	0.783	0.270
DERM7PT-C	827	2.282	15,442	16,318	0.946	0.086
DERM7PT-D	827	2.282	15,971	16,866	0.947	0.087
HAM10000	7,818	6.024	8,705	9,770	0.891	0.213
ISBI2016	1,273	4.092	10,553	10,992	0.960	0.101
ISBI2017	2,745	4.259	9,280	9,674	0.959	0.089
MED-NODE	170	1.429	9,029	9,513	0.949	0.068
MSK-1	1,088	2.615	11,753	14,068	0.835	0.173
MSK-2	1,522	3.299	9,288	9,418	0.986	0.062
MSK-3	225	10.842	8,075	8,074	1.000	0.112
MSK-4	943	3.366	6,930	7,162	0.968	0.065
PH2	200	4.000	12,688	14,928	0.850	0.210
SDC-198	648	4.735	14,054	14,840	0.947	0.116
UDA-1	557	2.503	11,730	12,243	0.958	0.083
UDA-2	60	1.609	11,297	11,601	0.974	0.020

3.1. Datasets

To validate the proposal, non-dermoscopic and dermoscopic images were obtained from several reputable sources, and can be consulted at the KDIS Research Group web page[2]. Table 1 shows a summary of the benchmark datasets. Only the images labeled as melanoma and nevus were considered, being in total 36,703 images. Most datasets present high imbalance ratio (ImbR), up to 10 in the case of MSK-3, which hampers the learning process. On the other hand, the intra-class (IntraC) and inter-class (InterC) metrics show the average distances between images belonging to different classes, as well as between images belonging to the same class. Both metrics were computed using the Euclidean function distance; each image i was represented as a vector. Then, the ratio (DistR) between these metrics showed that both distances are similar, which commonly indicates a high degree of overlapping between classes. Finally, the silhouette score (Silho) [39] was calculated, representing how similar an image is to its own cluster compared to other clusters. The results indicated that images were not well matched to their own cluster, and even samples belonging to different clusters are close in the feature space.

3.2. Experimental settings

Firstly, we trained the proposed architecture over each problem domain for 400 epochs. The goal was to obtain generators capable of create high-quality and realistic images. As mentioned before, generator and discriminator models are trained to maintain an equilibrium. In consequence, there are not many objective loss

[2] http://www.uco.es/kdis/skin-diagnosis-pggan/

Table 2: Basic configuration used.

Parameter	Value
Segmentation threshold	40%
Number of epochs (U-Net, R2U-Net)	150
M_g	400
Rotations	[1°,270°]
Flip	vertical and horizontal
Translations in X and Y	[-30%,30%]
Crop	[10%,30%]
Number of epochs	150
Mini-batch size	8
Learning rate (α)	SGD=0.01

functions used to effectively train both generator and discriminator. In this regard, Wasserstein loss was used to train both discriminator and generator architectures. The above method obtains a training process more stable and less sensitive to the architecture and hyperparameters [37]. Also, this loss is related with the quality of the images, showing properties of convergence. Although this means that it is not necessary to evaluate the generated samples looking for failures, we manually compare them with the real ones in order to assess the quality of the generator [40], which is a common practice. In addition, the proposed architecture is objectively evaluated by applying the Inception Score (IS), which seeks to assess the image quality and diversity [41]. This is perhaps the most widely adopted score for GAN evaluation. On the other hand, it was used a linear activation function in the output layer of the discriminator model, instead of sigmoid. Also, Karras et al. [34] recommended using Adam as optimization algorithm with a small learning rate ($\alpha = 0.001$, $\beta_1 = 0$, $\beta_2 = 0.99$), and as well as low momentum.

Secondly, InceptionV3, DenseNet, MobileNet, Xception and NASNetMobile convolutional architectures were assessed by using the synthetic images generated with the proposed architecture. The above results were compared to the same CNN models, but training with other classical and modern data augmentation techniques, such as random data augmentation, RE, UDA and a standard Progressive Generative Adversarial Networks. In this manner, we quantitatively evaluate which data augmentation technique obtains the higher predictive performance in the diagnosis of melanoma, which is the main aim of this work. Random Erasing selects a rectangle region in an image and erases its pixels. Training images with various levels of occlusion reduces the risk of overfitting and makes the model robust. The above can be integrated with most of the CNN-based recognition models, such as InceptionV3. On the other hand, Unsupervised Data Augmentation represents a new perspective on how to effectively noise unlabeled examples and supports that advanced data augmentation methods plays a crucial role in semi-supervised learning. This method consists in substituting simple noising operations with advanced data augmentation methods such as RandAugment [31] and back-translation [32].

Thirdly, each tuple CNN and data augmentation technique was evaluated by using non-segmented and segmented images in order to discover if the preprocessing step is suitable. To be fair, transfer learning was applied in all cases. Table 2 shows the configuration used to train all the models: the learning rate (α) was equal to 0.01 and it was reduced by a factor of 0.2 if an improvement in predictive performance was not observed during 10 epochs; a batch of size 8 was used due the medium size of the datasets; and the models were trained along 150 epochs. *Stochastic Gradient Descent* (SGD) [42] was used for training the models. SGD is one of the most used optimizers for training CNN models and despite its simplicity, it performs well across a variety of applications [43] and has been successfully applied for training networks in melanoma diagnosis [6, 14]. Regarding the tuning of the hyper-parameters of SGD, it is noteworthy that finding the optimal set of the hyper-parameter values is a task that commonly requires expensive and arduous work due to the many possible combinations [44]. In this work, a tuning process was not carried out and so the results could not be conferred to an over-adjustment. The datasets utilized in this work correspond to binary classification problems, so the cost function used for training the models was defined as the average of the binary cross entropy along all training samples. Data augmentation techniques were applied only in training data and test data were left untouched. The generated training images were considered as independent from the original ones.

3.3. Evaluation process

Regarding the evaluation process of the GANs architectures, a manual inspection of the generated images were performed and the IS values were assessed. The IS has a lowest value of 1 and a highest value of the number of classes supported by the classification model; in our case, the highest score is 2. The higher the IS, the better the image quality. The IS is computed as the Kullback-Leibler (KL) divergence summed over all images and averaged over the two classes:

$$\text{IS} = exp(E_x KL(p(y|x)||p(y))). \tag{7}$$

In order to assess the CNN models when training with the augmented data, a 3-times 10-fold cross validation process was performed on the datasets, and the results were averaged across all fold executions. In each fold, *Matthews Correlation Coefficient* (MCC) was used to measure the predictive performance of the models. MCC is widely used in Bioinformatics as performance metric [6, 15, 45], and it is specially designed to analyze the predictive performance on unbalanced data, which is common in skin lesion datasets (Table 1). MCC is computed as:

$$\text{MCC} = \frac{t_p \times t_n - f_p \times f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}}, \tag{8}$$

where t_p, t_n, f_p, and f_n are the number of true positives, true negative, false positives, and false negatives, respectively. MCC value is always in range $[-1, 1]$, where 1 represents a perfect prediction, 0 indicates a performance similar to a random prediction, and -1 an inverse prediction.

Finally, non-parametric statistical tests were used to detect whether there was any significant difference in predictive performance. Wilcoxon Signed-Rank [46] was performed when only two methods were compared. Friedman's test [47] was conducted in cases where a multiple comparison was carried out. After that, Hommel's test [48] was applied to detect significant differences with a control algorithm. All hypothesis testing were conducted at 95% confidence.

3.4. Software and hardware

The experimental study was executed with Ubuntu 18.04, four GPUs NVIDIA Geforce RTX 2080-Ti and four GPUs NVIDIA Geforce RTX 1080-Ti. All the experiments were implemented in Python v3.6, and the CNN models were developed by using Keras and TensorFlow as backend.

3.5. Results and discussions

In this section the results are presented. First, we analyzed qualitatively and quantitatively the images generated by the proposal. Second, the performance when training five CNN models with the generated images was assessed. Finally, the proposal was compared to several state-of-the-art data augmentation techniques.

3.5.1. Evaluating Generative Adversarial Networks

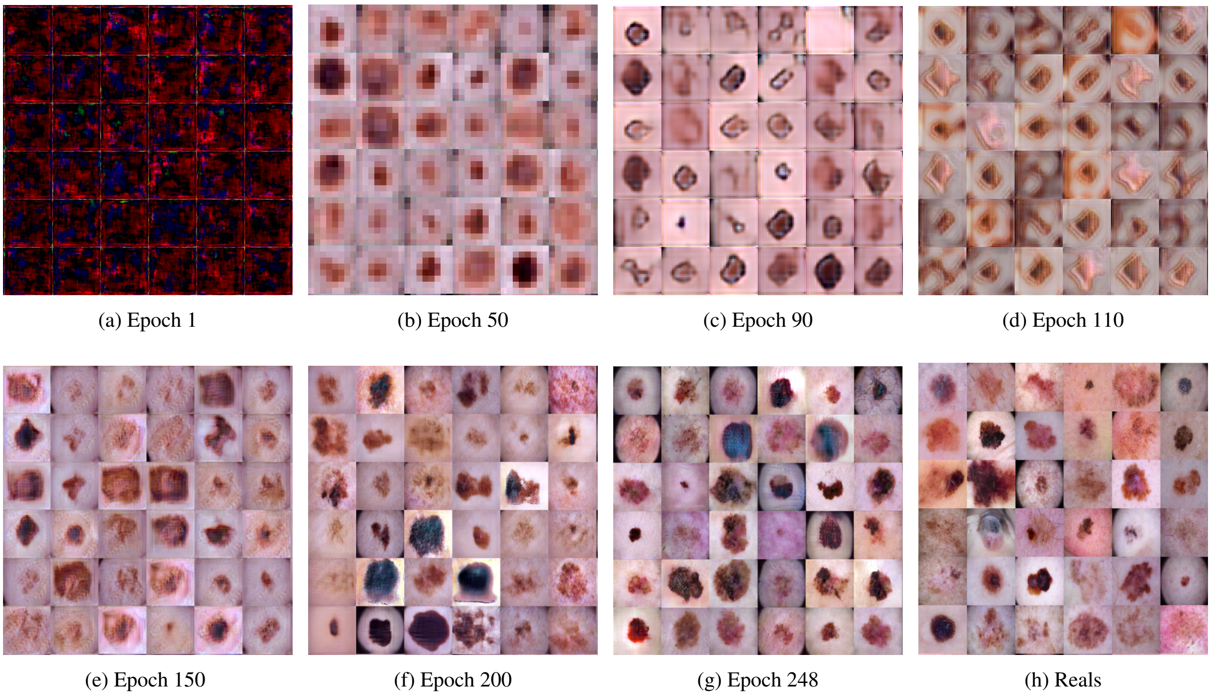
Figure 4 shows images generated by the proposal during several epochs. In this scenario, it took more than 200 epochs to notice some real progress in the synthetic images. We qualitatively assessed the images augmented by our proposal, where an increase in the stability of the models was observed, particularly in small datasets. In addition, plausible images were found in a lower number of epochs in some datasets (≈ 200 epochs). It is hard to point out differences between Figure 4g and Figure 4h regarding quality. In overall, after a high number of epochs, the generators mimic the real data very well, obtaining high-quality images.

Table 3: Average IS values of the baseline progressive growing of GAN and the proposal RGAN; "%" represents the fold changes after comparing the proposed model with the baseline. Wilcoxon's test rejected the null hypothesis with a p-value equal to 2.189E-4.

Dataset	GAN	RGAN	%
BCN20000	1.885	**1.992**	6
DERM-LIB	1.876	**1.986**	6
DERM7PT-C	1.866	**1.993**	7
DERM7PT-D	1.873	**1.993**	6
HAM10000	1.882	**1.996**	6
ISBI2016	1.758	**1.948**	11
ISBI2017	1.546	**1.700**	10
MED-NODE	1.841	**1.961**	7
MSK-1	1.873	**1.990**	6
MSK-2	1.871	**1.989**	6
MSK-3	1.750	**1.947**	11
MSK-4	1.856	**1.978**	7
PH2	1.856	**1.980**	7
SDC-198	1.867	**1.988**	7
UDA-1	1.857	**1.990**	7
UDA-2	1.782	**1.897**	6

On the other hand, Figure 5 and 6 show a quantitatively progressive evaluation using the IS. The IS required square images of about 300×300 pixels and an equal number of images in each category. The above requirements were fulfilled since 512×512 images were used and the same number of images peer category for evaluating the proposal. The attention was focused during 200-250 epochs, where a suitable performance was achieved in overall. It is noteworthy that the learning curves of the proposal were more stable compared to the standard GANs for diagnosing melanoma. HAM10000, DERM7PT-C and DERM7PT-D were the more feasible data sources, where the proposal achieved the top average performance. The biggest improvements were detected in ISBI2016 and MSK-3, which could be related with better predictive performance when training the CNN models in the next phase. Table 3 summarizes the top performance in each dataset, where our proposal achieved the highest result all the time, confirming the benefit and effectiveness of using the proposed double residual architecture for generating plausible images.

Although the IS indicated high-quality generated images, it is necessary to double-check this through the diagnostic capability of CNN models that are trained with the generated images. As a result, in the next sec-



(a) Epoch 1

(b) Epoch 50

(c) Epoch 90

(d) Epoch 110

(e) Epoch 150

(f) Epoch 200

(g) Epoch 248

(h) Reals

Figure 4: Manual inspection of the images generated by using the proposed double residual architecture.

tion our proposal and state-of-the-art data augmentation techniques are evaluated by using several CNN models.

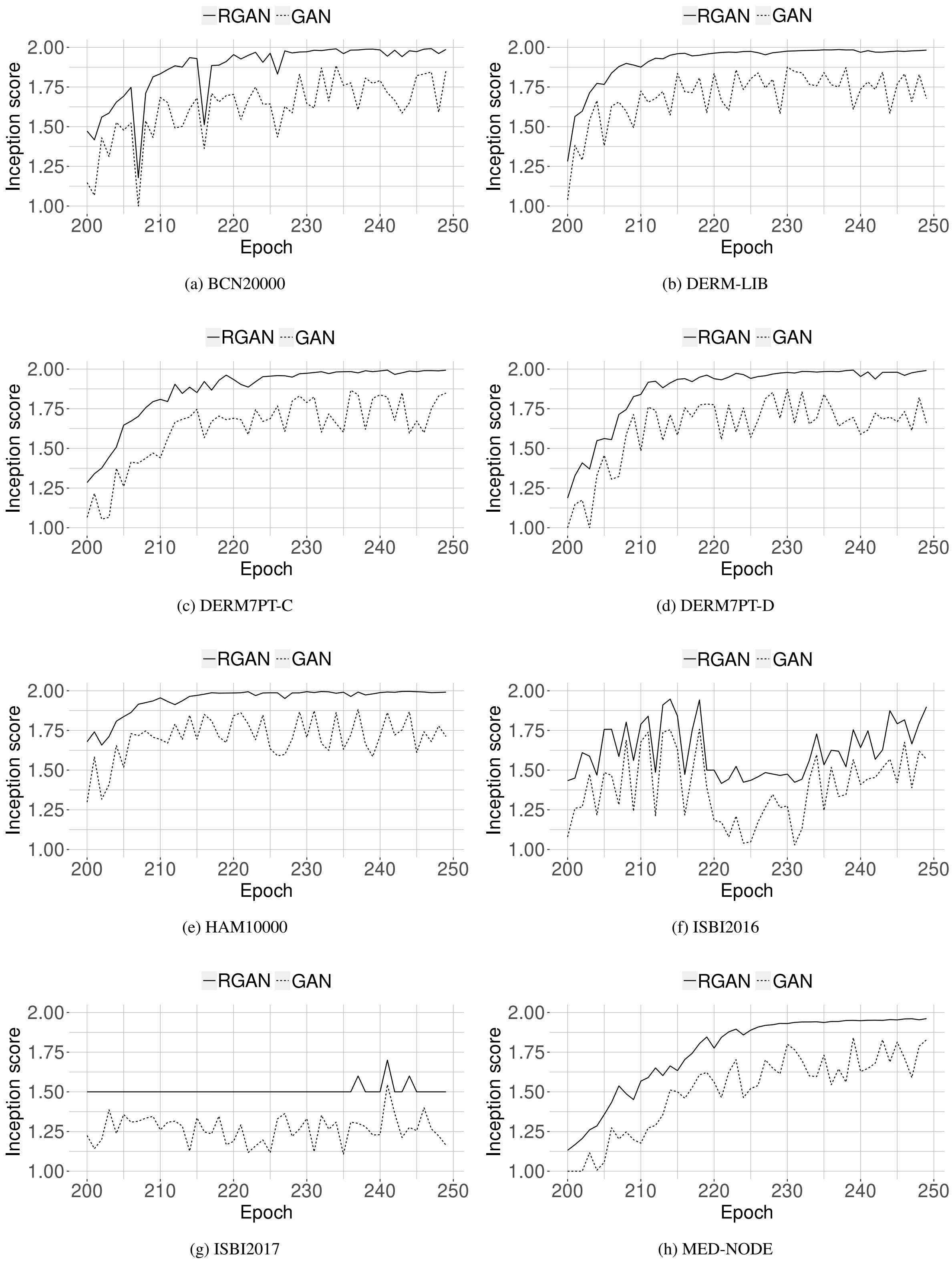
3.5.2. *Comparing with state-of-the-art techniques*

Table 4-8 show each CNN model trained using basic and advanced data augmentation techniques and the proposed method. The best MCC value by dataset was highlighted in bold typeface. As can be seen, the segmentation method helped the CNN models to achieve better performance, e.g. InceptionV3, RDA and segmented images outperformed its baseline by 12% in PH2 dataset.

Table 4 shows the results of NASNet model, where the proposal achieved the best performance all the time, except in BCN20000 and SDC-198 with non-segmented images, and only in SDC-198 when using segmented images. It is noteworthy that the proposal achieved 201% and 120% better performance compared to the average state-of-the-art techniques in MSK-3. The Friedman's test rejected the null hypothesis with a p-value equal to 2.870E-6; Friedman's statistic was equal to 31.138 with four degrees of freedom. The Friedman's ranking shows that the proposal obtained the first position, indicating that this model in average achieved a better performance than the rest of methods. Afterwards, the Hommel's post-hoc test was conducted by considering the proposal as the control method, and the results showed the proposal significantly outperformed the rest of the state-of-the-art techniques.

Table 5 shows the results of DenseNet201, where the proposal achieved the best performance the 81% and 75% of the time using non-segmented and segmented images, respectively. It is noteworthy that the proposal achieved 156% and 123% better performance compared to the average state-of-the-art techniques in MSK-3. The Friedman's test rejected the null hypothesis with a p-value equal to 1.023E-4; Friedman's statistic was equal to 23.463 with four degrees of freedom. The Friedman's ranking shows that the proposal obtained the first position, indicating that this model in average achieved a better performance than the rest of methods. Unsupervised data augmentation technique achieved the second best performance. Afterwards, the Hommel's post-hoc test was conducted by considering the proposal as the control method, and the results showed the proposal significantly outperformed the rest of the state-of-the-art techniques.

Table 6 shows the results of InceptionV3, where the proposal achieved the best performance the 75% and 81% of the time using non-segmented and segmented



(a) BCN20000

(b) DERM-LIB

(c) DERM7PT-C

(d) DERM7PT-D

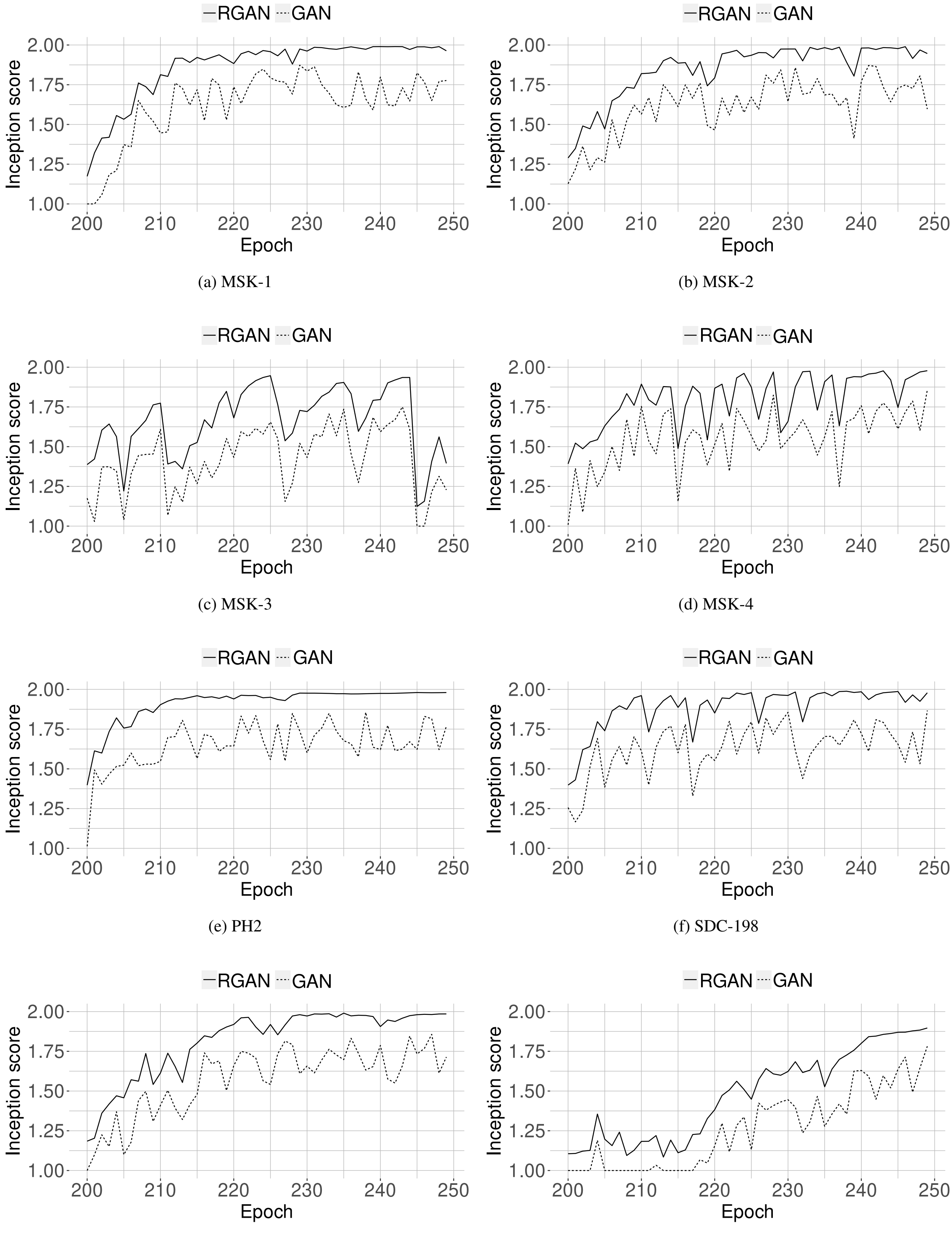
(e) HAM10000

(f) ISBI2016

(g) ISBI2017

(h) MED-NODE

Figure 5: Inception score. “GAN” and “RGAN” mean the default Progressive Growing of Generative Adversarial Networks and the proposed architecture, respectively.



(a) MSK-1

(b) MSK-2

(c) MSK-3

(d) MSK-4

(e) PH2

(f) SDC-198

(g) UDA-1

(h) UDA-2

Figure 6: Inception score. “GAN” and “RGAN” mean the default Progressive Growing of Generative Adversarial Networks and the proposed architecture, respectively.

Table 4: Average MCC values obtained by using NASNet. “RDA” means random data augmentation; “RE” and “UA” represent Random Erasing and Unsupervised Data Augmentation, respectively; “GAN” and “RGAN” mean the default progressive growing architecture and our proposal, respectively; “Avg” means the average performance of all the baseline methods and the “%” columns represent the fold changes when comparing the proposed model with the “Avg” column; “Ranking” means the average ranking computed by Friedman’s test and Hommel’s p-values are showed in the last row. The Friedman’s test rejected the null hypothesis with a p-value equal to 2.870E-6 when not using segmented images; Friedman’s statistic was equal to 31.138 with four degrees of freedom. The Friedman’s test rejected the null hypothesis with a p-value equal to 6.954E-6 when using segmented images; Friedman’s statistic was equal to 29.25 with four degrees of freedom.

Dataset	Non-segmented							Segmented						
	RDA	RE	UA	GAN	Avg	RGAN	%	RDA	RE	UA	GAN	Avg	RGAN	%
BCN20000	**0.749**	0.659	0.664	0.658	0.682	0.747	9	0.754	0.659	0.700	0.688	0.700	**0.780**	11
DERM-LIB	0.886	0.900	0.900	0.922	0.902	**0.986**	9	0.975	0.930	0.906	0.942	0.938	**1.000**	7
DERM7PT-C	0.471	0.330	0.474	0.419	0.424	**0.529**	25	0.500	0.350	0.485	0.459	0.448	**0.545**	22
DERM7PT-D	0.617	0.552	0.551	0.511	0.558	**0.698**	25	0.657	0.590	0.570	0.531	0.587	**0.736**	25
HAM10000	0.702	0.696	0.702	0.712	0.703	**0.781**	11	0.759	0.707	0.734	0.762	0.740	**0.781**	5
ISBI2016	0.429	0.285	0.254	0.349	0.329	**0.443**	35	0.471	0.325	0.274	0.379	0.362	**0.488**	35
ISBI2017	0.447	0.421	0.409	0.415	0.423	**0.505**	19	0.487	0.432	0.417	0.455	0.448	**0.512**	14
MED-NODE	0.633	0.787	0.887	0.700	0.752	**0.997**	33	0.673	0.826	0.925	0.710	0.784	**1.000**	28
MSK-1	0.669	0.701	0.716	0.750	0.709	**0.827**	17	0.704	0.707	0.737	0.760	0.727	**0.838**	15
MSK-2	0.423	0.439	0.463	0.446	0.443	**0.550**	24	0.459	0.475	0.482	0.476	0.473	**0.558**	18
MSK-3	0.232	0.225	0.452	0.000	0.227	**0.685**	201	0.498	0.238	0.481	0.040	0.314	**0.691**	120
MSK-4	0.481	0.593	0.618	0.552	0.561	**0.675**	20	0.490	0.626	0.627	0.592	0.584	**0.684**	17
PH2	0.741	0.900	0.900	0.900	0.860	**0.988**	15	0.856	0.915	0.925	0.940	0.909	**1.000**	10
SDC-198	0.603	0.617	**0.788**	0.598	0.652	0.649	0	0.614	0.654	**0.822**	0.618	0.677	0.651	-4
UDA-1	0.539	0.537	0.584	0.495	0.539	**0.667**	24	0.540	0.556	0.611	0.545	0.563	**0.682**	21
UDA-2	0.472	0.900	0.900	0.900	0.793	**0.956**	21	0.459	0.920	0.908	0.950	0.809	**1.000**	24
Ranking	3.656	3.656	2.875	3.688		**1.125**		3.563	3.750	3.188	3.375		**1.125**	
p-values	1.191E-5	1.191E-5	1.745E-3	9.126E-6		-		3.896E-5	1.063E-5	2.247E-4	1.140E-4		-	

Table 5: Average MCC values obtained by using DenseNet201. The Friedman’s test rejected the null hypothesis with a p-value equal to 1.023E-4 when not using segmented images; Friedman’s statistic was equal to 23.463 with four degrees of freedom. The Friedman’s test rejected the null hypothesis with a p-value equal to 7.904E-4 when using segmented images; Friedman’s statistic was equal to 18.988 with four degrees of freedom.

Dataset	Non-segmented							Segmented						
	RDA	RE	UA	GAN	Avg	RGAN	%	RDA	RE	UA	GAN	Avg	RGAN	%
BCN20000	0.771	0.687	0.690	0.698	0.712	**0.786**	10	0.792	0.716	0.701	0.708	0.729	**0.795**	9
DERM-LIB	0.966	0.900	0.900	0.900	0.916	**0.967**	6	0.992	0.937	0.923	0.910	0.941	**1.000**	6
DERM7PT-C	0.535	0.517	0.400	0.513	0.491	**0.590**	20	0.550	0.532	0.439	0.523	0.511	**0.624**	22
DERM7PT-D	0.655	0.533	0.689	0.547	0.606	**0.767**	27	0.636	0.564	0.723	0.547	0.618	**0.767**	24
HAM10000	0.747	0.726	0.708	0.707	0.722	**0.748**	4	0.698	0.738	0.729	0.737	0.726	**0.756**	4
ISBI2016	0.437	0.344	0.356	0.350	0.372	**0.441**	19	0.447	0.357	0.384	0.350	0.385	**0.450**	17
ISBI2017	0.447	0.428	0.429	0.393	0.424	**0.517**	22	0.419	0.437	0.449	0.413	0.430	**0.517**	20
MED-NODE	0.694	0.689	0.700	0.787	0.718	**0.878**	22	0.743	0.708	0.732	0.787	0.742	**0.883**	19
MSK-1	0.696	0.762	**0.807**	0.741	0.752	0.758	1	0.760	0.770	**0.841**	0.771	0.786	0.791	1
MSK-2	0.442	0.486	0.434	**0.502**	0.466	0.453	-3	0.436	0.521	0.472	**0.552**	0.495	0.443	-11
MSK-3	0.232	0.000	0.591	0.222	0.261	**0.670**	156	0.390	0.006	0.622	0.222	0.310	**0.691**	123
MSK-4	0.564	**0.625**	0.590	0.555	0.584	0.570	-2	0.533	**0.654**	0.620	0.605	0.603	0.565	-6
PH2	0.815	0.900	0.900	0.900	0.879	**0.981**	12	0.913	0.921	0.918	0.940	0.923	**1.000**	8
SDC-198	0.662	0.480	0.607	0.605	0.588	**0.699**	19	0.690	0.495	0.611	0.605	0.600	**0.707**	18
UDA-1	0.555	0.637	0.637	0.592	0.605	**0.644**	6	0.594	0.653	**0.668**	0.632	0.637	0.653	3
UDA-2	0.424	0.347	0.707	0.607	0.521	**0.974**	87	0.646	0.369	0.723	0.627	0.591	**1.000**	69
Ranking	3.188	3.719	3.031	3.688		**1.375**		3.500	3.406	2.938	3.625		**1.531**	
p-values	2.371E-3	8.272E-5	3.049E-3	1.057E-4		-		1.194E-3	1.592E-3	1.188E-2	7.204E-4		-	

Table 6: Average MCC values obtained by using InceptionV3. The Friedman's test rejected the null hypothesis with a p-value equal to 7.288E-6 when not using segmented images; Friedman's statistic was equal to 29.15 with four degrees of freedom. The Friedman's test rejected the null hypothesis with a p-value equal to 4.785E-7 when using segmented images; Friedman's statistic was equal to 34.938 with four degrees of freedom.

Dataset	Non-segmented							Segmented						
	RDA	RE	UA	GAN	Avg	RGAN	%	RDA	RE	UA	GAN	Avg	RGAN	%
BCN20000	0.745	0.678	0.667	0.668	0.690	**0.758**	10	0.785	0.699	0.684	0.718	0.722	**0.788**	9
DERM-LIB	0.954	0.900	0.900	0.922	0.919	**0.975**	6	0.968	0.031	0.004	0.922	0.481	**1.000**	108
DERM7PT-C	0.499	0.349	0.415	0.418	0.420	**0.548**	30	0.507	0.362	0.446	0.418	0.433	**0.584**	35
DERM7PT-D	**0.622**	0.478	0.489	0.534	0.531	0.594	12	0.636	0.494	0.528	0.534	0.548	**0.641**	17
HAM10000	0.684	0.675	0.632	0.634	0.656	**0.691**	5	0.651	0.711	0.655	0.674	0.673	**0.739**	10
ISBI2016	0.452	0.372	0.404	0.288	0.379	**0.485**	28	0.464	0.391	0.433	0.328	0.404	**0.517**	28
ISBI2017	0.416	0.290	0.294	0.302	0.326	**0.447**	37	0.392	0.329	0.311	0.312	0.336	**0.479**	43
MED-NODE	0.732	0.549	0.789	0.797	0.717	**0.842**	17	0.836	0.569	0.793	0.817	0.754	**0.887**	18
MSK-1	0.682	0.641	0.604	0.618	0.636	**0.779**	22	0.708	0.643	0.624	0.638	0.653	**0.812**	24
MSK-2	0.473	0.379	**0.487**	0.331	0.418	0.462	11	0.440	0.414	**0.517**	0.361	0.433	0.510	18
MSK-3	0.239	0.265	0.000	0.000	0.126	**0.659**	423	0.577	0.271	0.022	0.010	0.220	**0.691**	214
MSK-4	0.493	0.403	0.535	**0.559**	0.498	0.531	7	0.531	0.413	0.562	**0.579**	0.521	0.571	10
PH2	0.803	0.688	0.740	0.900	0.783	**0.996**	27	0.925	0.705	0.741	0.910	0.820	**1.000**	22
SDC-198	0.584	0.540	0.517	**0.604**	0.561	0.602	7	0.630	0.543	0.518	**0.654**	0.586	0.651	11
UDA-1	0.496	0.483	0.274	0.437	0.423	**0.695**	64	0.561	0.508	0.310	0.437	0.454	**0.700**	54
UDA-2	0.381	0.447	0.607	0.447	0.471	**0.993**	111	0.499	0.457	0.643	0.457	0.514	**1.000**	95
Ranking	2.563	4.000	3.813	3.250		**1.375**		2.500	3.969	3.875	3.469		**1.188**	
p-values	3.365E-2	1.063E-5	3.896E-5	1.592E-3		-		1.888E-2	2.607E-6	4.584E-6	8.975E-5		-	

Table 7: Average MCC values obtained by using MobileNet. The Friedman's test rejected the null hypothesis with a p-value equal to 3.940E-6 when not using segmented images; Friedman's statistic was equal to 30.463 with four degrees of freedom. The Friedman's test rejected the null hypothesis with a p-value equal to 1.741E-6 when using segmented images; Friedman's statistic was equal to 32.2 with four degrees of freedom.

Dataset	Non-segmented							Segmented						
	RDA	RE	UA	GAN	Avg	RGAN	%	RDA	RE	UA	GAN	Avg	RGAN	%
BCN20000	**0.732**	0.654	0.664	0.626	0.669	**0.732**	9	0.744	0.661	0.702	0.666	0.693	**0.759**	9
DERM-LIB	0.945	0.900	0.900	0.822	0.892	**0.959**	8	0.967	0.915	0.928	0.852	0.916	**1.000**	9
DERM7PT-C	0.490	0.413	0.517	0.445	0.466	**0.534**	15	0.546	0.437	0.533	0.485	0.500	**0.579**	16
DERM7PT-D	0.657	0.608	0.543	0.511	0.580	**0.682**	18	0.642	0.620	0.549	0.551	0.590	**0.690**	17
HAM10000	0.701	0.662	0.685	0.660	0.677	**0.726**	7	0.672	0.694	0.696	0.680	0.686	**0.756**	10
ISBI2016	0.422	0.300	0.256	0.361	0.335	**0.481**	44	0.454	0.336	0.256	0.391	0.359	**0.494**	38
ISBI2017	**0.431**	0.350	0.392	0.352	0.381	0.397	4	0.430	0.389	0.429	0.392	0.410	**0.433**	6
MED-NODE	0.717	0.672	0.449	0.789	0.657	**0.873**	33	0.768	0.697	0.477	0.839	0.695	**0.887**	28
MSK-1	0.700	**0.746**	0.716	0.693	0.714	0.743	4	0.725	0.764	0.735	0.743	0.742	**0.789**	6
MSK-2	0.472	0.486	0.466	0.433	0.464	**0.527**	14	0.401	0.520	0.477	0.473	0.468	**0.551**	18
MSK-3	0.294	0.104	**0.697**	0.000	0.274	0.661	141	0.468	0.121	**0.710**	0.050	0.337	0.691	105
MSK-4	0.525	0.525	0.446	0.520	0.504	**0.553**	10	0.502	0.528	0.479	0.570	0.520	**0.572**	10
PH2	0.818	0.900	0.900	0.840	0.864	**0.967**	12	0.937	0.926	0.930	0.870	0.916	**1.000**	9
SDC-198	0.653	0.601	0.540	0.604	0.600	**0.757**	26	0.667	0.636	0.570	0.604	0.619	**0.770**	24
UDA-1	0.551	0.592	0.589	0.584	0.579	**0.665**	15	0.599	0.600	0.605	0.584	0.597	**0.682**	14
UDA-2	0.375	0.607	0.900	0.707	0.647	**0.961**	48	0.658	0.642	0.923	0.757	0.745	**1.000**	34
Ranking	2.875	3.406	3.375	4.125		**1.219**		3.063	3.688	3.375	3.813		**1.063**	
p-values	3.049E-3	1.822E-4	2.294E-4	8.021E-7		-		3.466E-4	7.969E-6	7.046E-5	3.473E-6		-	

Table 8: Average MCC values obtained by using Xception. The Friedman's test rejected the null hypothesis with a p-value equal to 5.854E-4 when not using segmented images; Friedman's statistic was equal to 19.65 with four degrees of freedom. The Friedman's test rejected the null hypothesis with a p-value equal to 1.421E-4 when using segmented images; Friedman's statistic was equal to 22.75 with four degrees of freedom.

Dataset	Non-segmented							Segmented						
	RDA	RE	UA	GAN	Avg	RGAN	%	RDA	RE	UA	GAN	Avg	RGAN	%
BCN20000	0.752	0.663	0.653	0.707	0.694	**0.766**	10	0.764	0.681	0.670	0.717	0.708	**0.785**	11
DERM-LIB	0.910	0.822	0.822	0.900	0.863	**0.953**	10	0.976	0.860	0.836	0.910	0.896	**1.000**	12
DERM7PT-C	0.484	0.419	0.392	0.451	0.436	**0.513**	18	0.500	0.423	0.413	0.451	0.447	**0.517**	16
DERM7PT-D	0.634	0.580	0.605	0.584	0.601	**0.753**	25	0.689	0.615	0.642	0.584	0.632	**0.767**	21
HAM10000	**0.691**	0.650	0.609	0.653	0.651	0.667	2	0.707	0.688	0.615	0.663	0.668	**0.713**	7
ISBI2016	0.421	0.411	0.354	0.254	0.360	**0.500**	39	0.488	0.424	0.358	0.284	0.388	**0.514**	32
ISBI2017	**0.402**	0.306	0.371	0.352	0.358	0.397	11	0.437	0.319	0.376	0.382	0.379	**0.439**	16
MED-NODE	0.717	0.789	0.789	0.789	0.771	**0.884**	15	0.734	0.820	0.822	0.839	0.804	**0.887**	10
MSK-1	0.665	0.721	0.700	0.659	0.686	**0.781**	14	0.699	0.721	0.732	0.679	0.708	**0.788**	11
MSK-2	0.421	0.473	0.493	**0.564**	0.488	0.455	-7	0.443	0.497	0.518	**0.604**	0.516	0.455	-12
MSK-3	0.195	0.069	0.041	0.000	0.076	**0.362**	375	0.324	0.071	0.058	0.050	0.126	**0.365**	190
MSK-4	0.459	0.525	0.453	0.507	0.486	**0.552**	14	0.487	0.527	0.463	0.537	0.504	**0.581**	15
PH2	0.727	0.766	**0.900**	0.800	0.798	0.821	3	0.854	0.790	**0.906**	0.810	0.840	0.866	3
SDC-198	0.640	0.607	0.601	0.532	0.595	**0.667**	12	0.661	0.616	0.620	0.572	0.617	**0.707**	15
UDA-1	0.504	**0.582**	**0.582**	0.503	0.543	0.557	3	0.525	0.605	**0.607**	0.553	0.572	0.580	1
UDA-2	0.307	0.900	0.707	0.333	0.562	**0.982**	75	0.461	0.928	0.743	0.343	0.619	**1.000**	62
Ranking	3.000	3.313	3.500	3.688		**1.500**		3.000	3.500	3.438	3.688		**1.375**	
p-values	7.290E-3	2.371E-3	1.040E-3	3.644E-4		-		3.650E-3	3.370E-4	4.494E-4	1.409E-4		-	

images, respectively. Once again, the proposal was slightly surpassed in MSK-2, MSK-4 and SDC-198, which stated these datasets as the most difficult to the proposal. It is noteworthy that the proposal achieved 423% and 214% better performance compared to the average state-of-the-art techniques in MSK-3. The Friedman's test rejected the null hypothesis with a *p*-value equal to 7.288E-6; Friedman's statistic was equal to 29.15 with four degrees of freedom. The Friedman's ranking shows that the proposal obtained the first position, indicating that this model in average achieved a better performance than the rest of methods. Random data augmentation technique achieved the second best performance. Afterwards, the Hommel's post-hoc test was conducted by considering the proposal as the control method, and the results showed the proposal significantly outperformed the rest of the state-of-the-art techniques.

Table 7 shows the results of MobileNet, where the proposal achieved the best performance all the time when using segmented images, except in MSK-3. The proposal was only slightly surpassed in 2.7% by UDA method. However, it is noteworthy that the proposal achieved 141% and 105% better performance compared to the average state-of-the-art techniques in MSK-3. The Friedman's test rejected the null hypothesis with a *p*-value equal to 3.940E-6; Friedman's statistic was equal to 30.463 with four degrees of freedom. The Friedman's ranking shows that the proposal obtained the first position, indicating that this model in average achieved a better performance than the rest of methods. Random data augmentation technique achieved the second best performance. Afterwards, the Hommel's post-hoc test was conducted by considering the proposal as the control method, and the results showed the proposal significantly outperformed the rest of the state-of-the-art techniques.

Table 8 shows the results of Xception, where the proposal did not achieve the best performance in HAM10000, ISBI2017, MSK-2, PH2 and UDA-1. However, it is noteworthy that the proposal achieved 375% and 190% better performance compared to the average state-of-the-art techniques in MSK-3. The Friedman's test rejected the null hypothesis with a *p*-value equal to 5.854E-4; Friedman's statistic was equal to 19.65 with four degrees of freedom. The Friedman's ranking shows that the proposal obtained the first position, indicating that this model in average achieved a better performance than the rest of methods. Random data augmentation technique achieved the second best performance. Afterwards, the Hommel's post-hoc test was conducted by considering the proposal as the control method, and the results showed the proposal significantly outperformed the rest of the state-of-the-art techniques.

Table 9: Average MCC values and parameters of the used CNN models. The models are sorted by average predictive performance.

Model	MCC	Parameters
NASNetMobile	**0.747**	5,326,716
MobileNet	0.728	**4,253,864**
DenseNet201	0.728	20,242,984
InceptionV3	0.723	23,851,784
Xception	0.685	22,910,480

Overall, the most suitable datasets were DERM-LIB, PH2 and MED-NODE with 89%, 88% and 77% MCC, respectively. On the other hand, the most complex datasets were MSK-3, ISBI2016 and ISBI2017 with 30%, 39% and 41% MCC, respectively. However, the proposal surpassed the average performance of its competitors by 190% when using segmented images and 375% when not in MSK-3. In addition, the proposal achieved always a 9% better performance compared to the average results in BCN20000, which is the largest dataset publicly available. Furthermore, the CNN models achieved the best average performance when training with the images generated by the proposal. The proposed RGAN achieved the best performance the 82% of the time.

To sum up, Table 9 shows the average performance of each CNN model and their number of trainable parameters, where NASNet achieved the top average performance and Xception ended last. These results indicate that a bigger number of trainable parameters will not necessarily obtain a better performance in melanoma diagnosis.

4. Conclusions

In this work, the diagnosis of melanoma was addressed via a series of contributions. Firstly, a double residual architecture was designed and applied on melanoma diagnosis in order to generate plausible synthetic skin images. The architecture was evaluated qualitatively and quantitatively by using a manual inspection and the IS score, respectively. These results showed stable learning even with a low number of samples. Then, an extensive experimental study was performed on sixteen skin image datasets. Overall, results showed that the proposed architecture significantly surpassed several state-of-the-art data augmentation techniques in five CNN models. The above corroborated the hypothesis that complex data augmentation techniques are suitable to train CNN models, even in small datasets with complex properties. In addition, to preprocess data, a segmentation method was applied. The results showed that all CNN models improved their average performance by using segmented data. The performance was increased by applying transfer learning from pre-trained ImageNet. Bear in mind that transfer learning alleviated the requirement for a large number of training data.

Future works will conduct more extensive experiments to validate the full potential of the proposed architecture, for example by considering a wide set of hyperparameters to be tuned. In addition, we look forward to implementing new evaluation metrics in order to maintain an equilibrium during the training process. Furthermore, it would be interesting to receive the feedback from dermatologists regarding realism. Finally, it is noteworthy that the proposed approach is not strictly restricted to melanoma diagnosis problem, and according the results it could be applied in the future on other complex real-world problems where data is limited.

Conflict of interest statement

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Spanish Ministry of Science and Innovation, the University of Cordoba, and the European Regional Development Fund, projects UCO-FEDER 18 REF.1263116 MOD and PID2020-115832GB-I00. It was also funded by the i-PFIS contract no. IFI17/00015 granted by the Health Institute Carlos III of Spain.

References

[1] American Cancer Society, Cancer Facts and Figures, consulted on June 22, 2021 (2021).
URL https://bit.ly/3gNDBVr

[2] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.

[3] H. Kittler, H. Pehamberger, K. Wolff, M. Binder, Diagnostic accuracy of dermoscopy, Lancet Oncology 3 (3) (2002) 159–165.

[4] A.-R. Ali, T. M. Deserno, A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data, in: Progress in Biomedical Optics and Imaging - Proceedings of SPIE, Vol. 8318, San Francisco, USA, 2012.

[5] J. Sánchez-Monedero, M. Pérez-Ortiz, A. Sáez, P. A. Gutiérrez, C. Hervás-Martínez, Partial order label decomposition approaches for melanoma diagnosis, Applied Soft Computing Journal 64 (2018) 341–355.

[6] E. Pérez, O. Reyes, S. Ventura, Convolutional neural networks for the automatic diagnosis of melanoma: An extensive experimental study, Medical Image Analysis 67.

[7] M. Binder, M. Schwarz, A. Winkler, A. Steiner, A. Kaider, K. Wolff, H. Pehamberger, Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists, Archives of dermatology 131 (3) (1995) 286–291.

[8] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. Soroushmehr, M. Jafari, K. Ward, K. Najarian, Melanoma detection by analysis of clinical images using convolutional neural network, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, Florida, USA, 2016, pp. 1373–1376.

[9] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task, European Journal of Cancer 111 (2019) 148–154.

[10] A. I. Khan, J. L. Shah, M. M. Bhat, CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images, Computer Methods and Programs in Biomedicine 196. doi:10.1016/j.cmpb.2020.105581.

[11] U. Asif, M. Bennamoun, F. Sohel, A Multi-Modal, Discriminative and Spatially Invariant CNN for RGB-D Object Labeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (9) (2018) 2051–2065.

[12] Ericsson, On the pulse of the networked society, Tech. rep. (2015).
URL https://apo.org.au/node/59109

[13] K. Lenc, A. Vedaldi, Understanding Image Representations by Measuring Their Equivariance and Equivalence, International Journal of Computer Vision 127 (5) (2019) 456–476.

[14] F. Perez, C. Vasconcelos, S. Avila, E. Valle, Data augmentation for skin lesion analysis, in: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, Springer, Granada, Spain, 2018, pp. 303–311.

[15] E. Pérez, S. Ventura, An ensemble-based convolutional neural network model powered by a genetic algorithm for melanoma diagnosis, Neural Computing and Applicationsdoi:10.1007/s00521-021-06655-7.

[16] J. Kawahara, S. Daneshvar, G. Argenziano, G. Hamarneh, Seven-point checklist and skin lesion classification using multitask multimodal neural nets, IEEE Journal of Biomedical and Health Informatics 23 (2) (2019) 538–546.

[17] C. Baur, S. Albarqouni, N. Navab, MelanoGANs: high resolution skin lesion synthesis with GANs, arXiv preprint arXiv:1804.04338.

[18] E. Pérez, S. Ventura, Melanoma recognition by fusing convolutional blocks and dynamic routing between capsules, Cancers 13 (19). doi:10.3390/cancers13194974.

[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 770–778.

[20] G. Huang, Z. Liu, L. Van Der Maaten, K. Weinberger, Densely connected convolutional networks, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 2017.

[21] D. C. Ciresan, U. Meier, L. M. Gambardella, J. Schmidhuber,

Deep, big, simple neural nets for handwritten digit recognition, Neural Computation 22 (12) (2010) 3207–3220.

[22] R. A. Van den Berg et al., Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC genomics 7 (1) (2006) 142.

[23] M. Al-masni, D.-H. Kim, T.-S. Kim, Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification, Computer Methods and Programs in Biomedicine 190. doi:10.1016/j.cmpb.2020.105351.

[24] M. Rubin, O. Stein, N. A. Turko, Y. Nygate, D. Roitshtain, L. Karako, I. Barnea, R. Giryes, N. T. Shaked, TOP-GAN: Stain-free cancer cell classification using deep learning with a small training set, Medical Image Analysis 57 (2019) 176–185.

[25] G. Liang, L. Zheng, A transfer learning method with deep residual network for pediatric pneumonia diagnosis, Computer Methods and Programs in Biomedicine 187. doi:10.1016/j.cmpb.2019.06.023.

[26] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random Erasing Data AugmentationarXiv:1708.04896.
URL https://arxiv.org/abs/1708.04896

[27] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, Q. V. Le, Unsupervised Data Augmentation for Consistency TrainingarXiv:1904.12848.
URL https://arxiv.org/abs/1904.12848

[28] A. Radford, L. Metz, S. Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial NetworksarXiv:1511.06434.
URL https://arxiv.org/abs/1511.06434

[29] E. Denton, S. Chintala, A. Szlam, R. Fergus, Deep generative image models using a laplacian pyramid of adversarial networks, in: Advances in Neural Information Processing Systems, Vol. 2015-Janua, Montreal, Canada, 2015.

[30] Z. Qin, Z. Liu, P. Zhu, Y. Xue, A gan-based image synthesis method for skin lesion classification, Computer Methods and Programs in Biomedicine 195. doi:10.1016/j.cmpb.2020.105568.

[31] E. D. Cubuk, B. Zoph, J. Shlens, Q. V. R. Le, Practical data augmentation with no separate search, arXiv preprint arXiv:1909.13719.

[32] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, arXiv preprint arXiv:1511.06709.

[33] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, Vol. 3, Montreal, Quebec, Canada, 2014, pp. 2672–2680.

[34] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: 6th International Conference on Learning Representations, Vancouver, Canada, 2018.

[35] S. Chintala, E. Denton, M. Arjovsky, M. Mathieu, How to train a GAN? Tips and tricks to make GANs work (2016).

[36] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE international conference on computer vision, Venice, Italy, 2017, pp. 2794–2802.

[37] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, arXiv preprint arXiv:1701.07875.

[38] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, California, USA, 2017, pp. 5767–5777.

[39] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.

[40] A. Borji, Pros and cons of GAN evaluation measures, Computer Vision and Image Understanding 179 (2019) 41–65.

[41] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, Advances in neural information processing systems 29 (2016) 2234–2242.

[42] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, Vol. 1, MIT press Cambridge, 2016.

[43] A. Pontoriero et al., Automated data quality control in fdopa brain pet imaging using deep learning, Computer Methods and Programs in Biomedicine 208. doi:10.1016/j.cmpb.2021.106239.

[44] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (Feb) (2012) 281–305.

[45] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using matthews correlation coefficient metric, PloS one 12 (6) (2017) e0177678.

[46] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (6) (1945) 80–83.

[47] M. Friedman, A comparison of alternative tests of significance for the problem of *m* rankings, The Annals of Mathematical Statistics 11 (1) (1940) 86–92.

[48] G. Hommel, A stagewise rejective multiple test procedure based on a modified bonferroni test, Biometrika 75 (2) (1988) 383–386.

7

Conference publications

7.1. Discovering multimodal fusion architectures for multi-class skin lesion diagnosis

Title	Multi-view Deep Neural Networks for multiclass skin lesion diagnosis
Authors	E. Pérez, S. Ventura
Conference	IEEE International Conference on Omni-layer Intelligent systems
Year	2022
Editorial	IEEE
DOI	10.1109/COINS54846.2022.9854997

Multi-view Deep Neural Networks for multiclass skin lesion diagnosis

Eduardo Pérez
Andalusian Research Institute in Data Science
and Computational Intelligence, DaSCI.
Maimonides Biomedical Research Institute of Córdoba.
University of Córdoba
Córdoba, Spain
0000-0003-2343-2634

Sebastián Ventura
Andalusian Research Institute in Data Science
and Computational Intelligence, DaSCI.
Maimonides Biomedical Research Institute of Córdoba.
University of Córdoba
Córdoba, Spain
0000-0003-4216-6378

Abstract—**Early diagnosis is still the best method to face skin cancer. The diagnosis of skin lesions remains as a challenge for physicians and researchers. In the past few years, it has benefited from computer-aided diagnosis methods that successfully apply classic Machine Learning techniques and more recently Convolutional Neural Networks. This work is aimed at discovering architectures that best fuse clinical records and medical images for the diagnosis of skin lesions. As a result, a genetic algorithm is designed in order to select how to combine such information and the main details of the new architecture. The architecture is able to cope with multiple inputs and learn multiple outputs, proving flexibility by sharing network parameters, which implicitly mitigates the overfitting of the model. An extensive experimental study was conducted on the well-known ISIC2019 dataset, where the models were trained with a total of 72,106 images and meta-data, including the augmented images. The proposal outperformed the baseline state-of-the-art model while diagnosing from eight skin lesion categories. Furthermore, the discovered architecture achieved 85%, 94%, and 84% of recall score when diagnosing malignant lesions - melanoma, basal cell carcinoma, and squamous cell carcinoma, respectively. Finally, the results showed the suitability of the proposed genetic algorithm, which was able to automatically build a multimodal fusion architecture for the diagnosis of skin lesions.**

Index Terms—**skin lesion diagnosis, melanoma, multimodal data, multi-view, deep learning, genetic algorithm**

I. Introduction

THE incidence and mortality rates of skin cancer remains as a huge concern in many countries. Only in the Unites States, 7,650 deaths are expected in 2022 (5,080 men and 2,570 women), which represents 470 more deaths than 2021 [1]. Recent works have started to examine automated image analysis techniques, from handcrafted features and classical machine learning techniques [2], to advanced deep learning methods [3]. The first methods require the previous extraction of handcrafted descriptors, which is prone to errors. However, deep learning methods can automatically extract and learn high-level features [3], providing a higher robustness to the inter- and intra-class variability present in melanoma images [4]. In addition, the above techniques may even outperform dermatologists in terms of diagnosis, and could be used as a diagnostic support tool [5].

On the other hand, CNN models still present several issues that hamper their correct application in the diagnosis of melanoma. For instance, CNN models are sensitive to variations in viewpoints, changes in lighting conditions, occlusions and background clutter [6]. In addition, CNN models fit a wide variety of non-linear data points and several of the existing public melanoma datasets encompass only a few hundred or thousand of images. As a result, CNN models are led to overfit, therefore attaining a poor generalization capacity. Furthermore, although the majority of melanoma datasets are comprised of dermoscopic images, collect images taken with common digital cameras can boost the development of modern tools for a less expensive skin lesion diagnosis, thus reducing invasive treatments and the required economical resources. Nowadays, the rise of the digital era makes it possible not only to obtain images at a low cost, but also to obtain all sorts of additional information, such as symptomatologies and others. In this regard, there is an opportunity to create models that can merge several types of input data to satisfy the above requirements. Accordingly, a proposal for combining images and meta-data is presented.

To begin with, multi-view learning is a paradigm that has shown to be suitable for improving the generalization capacity of machine learning models [7]. Its main goal is to learn one function that represents both image and meta-data, and optimizes the architecture in order to improve the generalization performance. Early, joint, and late fusion are well-known and proven approaches to fuse the information from the views, which are used in the present work. Early fusion joins raw or preprocessed features from different modalities before passing them to a predictive model, which will obtain a single prediction [8]. On the other hand, joint fusion can use a model for each view, which will output abstract features. Such features are then combined and passed to a final predictive model. All models will be trained by propagating the final loss from

This research was funded by the Spanish Ministry of Science and Innovation, the University of Cordoba, and the European Regional Development Fund, projects PID2020-115832GB-I00 and UCO-FEDER 18 REF.1263116 MOD. It was also funded by the i-PFIS contract no. IFI17/00015 granted by Health Institute Carlos III of Spain.

978-1-6654-8356-8/22/$31.00 ©2022 IEEE

the predictive model [9]. In this manner, all models should be improved at the same time. Finally, late fusion constructs independent models and obtains their individual predictions [10]. After that, the final prediction is made by aggregating the above predictions using averaging, majority voting, among others. For example, a multi-view model for thyroid nodules diagnosis was developed by analyzing information from raw ultrasound images, manually extracted medical features from segmentation analysis, and statistical and texture features [11]. The results showed that the best performance was found by applying a late fusion approach and a majority vote on the classification results of the three views.

On the other hand, the inclusion of metadata on previous researchs suggests that the performance could be improved [9]. It makes sense that complex relationships exist between the characteristics of a sample, even more in human samples. So far, age, anatomical site and sex can be found in some relevant public skin image datasets, which makes possible to design advanced models. In addition, nowadays the tendency is to increase data collection, and then try to find complex relationships in the data. In the near future, it is highly possible that clinical background of patients and its evolution will be used in more complex models, fusing data from different sources.

To sum up, the multiview learning problem deals with the prediction of multiple variables from a set of input variables, in our case images and metadata; what is learned for each model can help in solving the overall prediction. In this research a new framework is proposed, which predicts which type of lesion a sample is. Learning from multiple inputs contributes to regularize the training process and implicitly mitigates the overfitting of the model. Aimed at finding how to efficiently merge both metadata and image, an algorithm is proposed in order to explore such challenge. In this work, it was hypothesised that different convolutional architectures would be cooperatively enhanced, if they share convolutional blocks that have proven to be effective in at least one of them. The main contributions from this manuscript are:

- A study of the utility of multimodal data and its inclusion within different architectures and training settings.
- A genetic algorithm, which is capable of discovering suitable architectures for the diagnosis of skin lesion.
- The proposed architecture is able to identify eight categories: (i) nevus, (ii) actinic keratosis, (iii) benign keratosis, (iv) dermatofibroma, (v) vascular lesion, (vi) melanoma, (vii) basal cell carcinoma, (viii) squamous cell carcinoma.

The rest of this work is arranged as follows: Section II gives details about the designed genetic algorithm, which was used to learn the suboptimal sets of architectures; Section III discusses the experimental study carried out, showing a discussion of them; finally, some concluding remarks are presented in Section IV.

II. Discovering multi-view architectures via a genetic algorithm

Regarding the generation of the different architectures, such models can be generated either applying randomness or finding a set of transformations by performing some type of guided search. As a consequence, a genetic algorithm (GA) was designed to learn the suboptimal architectures. Fig. 1 shows an overview of the three possible baseline architectures. We selected the most feasible, yet difficult, variables to evolve: feed forward settings, such as number of layers, units and output features ($f_1, f_2, f_i, ..., f_m$); CNN output features ($c_1, c_2, c_i, ..., c_n$); and the prediction block settings. An effective configuration for a block of the architecture in Fig. 1a could be also valid in Fig. 1b and Fig. 1c, and otherwise. Following subsections explain each one of the component of the designed GA.

A. Individuals and chromosome codification

Let say that the population of the GA has k individuals $\{I_1, I_2, \ldots, I_k\}$, where the j-th individual (I_j) has a chromosome encoded as a ordered list of values, as shown in Fig. 2. A gene g_p^j represents the p-th gene of the individual I_j, and denotes a setting to be applied on the architecture; e.g. the number of hidden layers in the feed forward network. Each chromosome, therefore, represents an ordered list of settings, and a full architecture. It should be noted that the length of a chromosome will be equal to the size of parameters to be tuned.

B. Creation of the initial population

In order to guarantee the diversity, the chromosome of each individual I_j of the population is randomly created, but repeated individuals are not allowed, avoiding the early convergence of the method to local minima. Two individuals are considered identical when all genes g_p^j in the chromosome are identical.

C. Fitness function

The fitness function measures the performance of the architecture, thus leading the evolution towards high-performing individuals. Many evaluation measures for multiclass prediction have been proposed in the literature, such as multiclass *Matthews Correlation Coefficient* (MCC). This metric is being used in several bioinformatic scenarios and has the property to summarize well the performance of the classifiers on complex data [12]. In addition, MCC is even more reliable than balanced accuracy in some scenarios [13], and is specially designed to analyze the predictive performance on unbalanced data, even if the classes are of very different sizes, which is present in this work. The fitness function used to evaluate the individual I_j can be calculated as

$$\uparrow f_j = MCC_j = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}}, \quad (1)$$

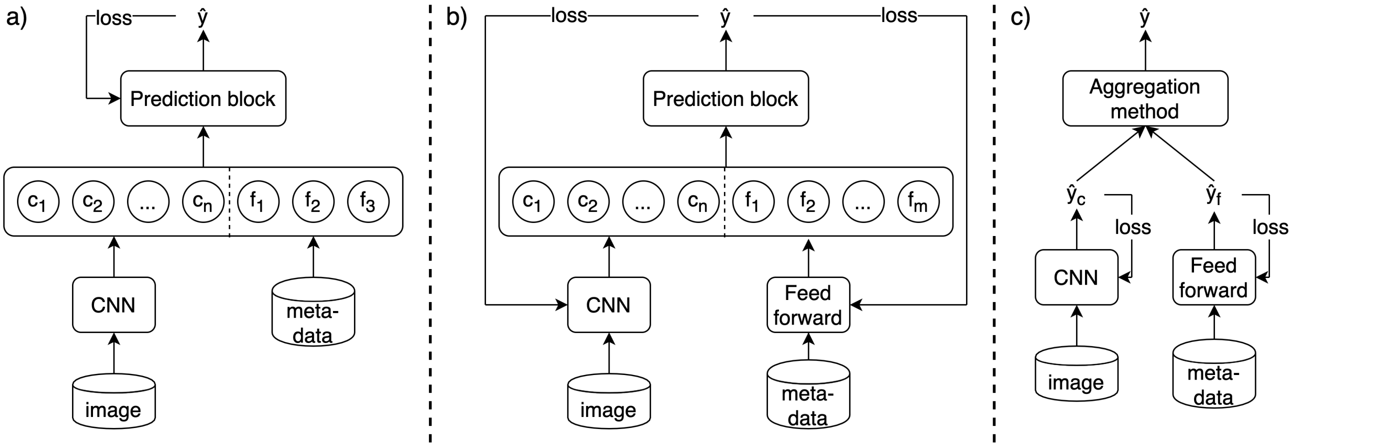


Fig. 1. Explored architectures by the proposed genetic algorithm; a) means early feature-level fusion; b) refers to joint training fusion; c) indicates independent decision fusion; $f_{\{1...h\}}$ are the meta-data feature representations; $c_{\{1...h\}}$ are the abstract feature representations extracted from images; $\hat{y}_c$ and $\hat{y}_f$ are partial predictions from the CNN and Feed Forward models, respectively; $\hat{y}$ is the final prediction. Given a sample, the model predicts its class by considering each particular case.

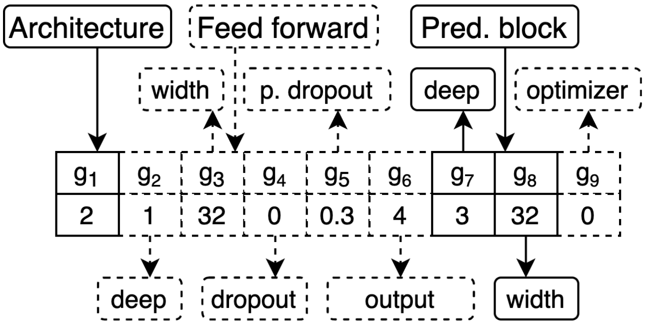


Fig. 2. Example of a chromosome. Figure shows how a new architecture is obtained by applying the ordered list of settings encoded by the chromosome.

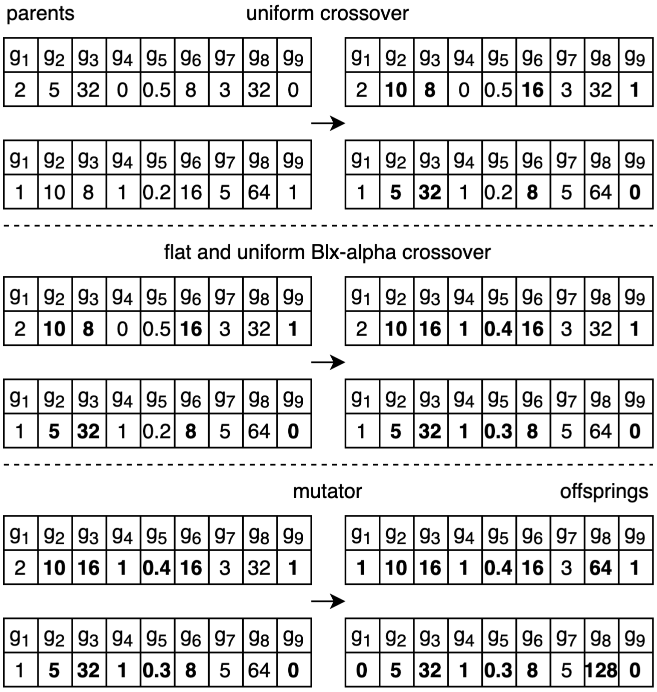
where ↑ means the fitness should be maximized; $t_k = \sum_i^K C_{ik}$ means the number of times class k truly occurred; $p_k = \sum_i^K C_{ki}$ represents the number of times class k was predicted; $c = \sum_k^K C_{kk}$ is the total number of samples correctly predicted; and $s = \sum_i^K \sum_j^K C_{ij}$ is the total number of samples.

D. Parent selection

A tournament size equal 2 was used in this work in order to lower the selection pressure and increase the search space. The parents are selected by a tournament selection procedure to create the intermediate population [14]. First, two individuals are randomly selected. Second, the individuals are compared and the best one is selected with replacement. Finally, the process is repeated and finishes when the number of individuals is completed.

E. Genetic operators

Fig. 3 shows the genetic operators applied. First, an Uniform crossover was carried out with a crossover rate p_{uc}. As such, the genes in the same locus k of the chromosomes j and q are swapped (g_k^j and g_k^q). Then, a custom Flat crossover for an integer representation was performed with a lower crossover rate p_{fc} [15]. As a result, a random integer value is chosen from the interval $[g_k^j, g_k^q]$. On the other hand, BLX-α crossover is applied in the gene that contains real numeric values - it chooses a uniform random real number from the interval $[g_5^j, g_5^q]$. Once the new offspring is generated, an one-point mutator operator is applied with a probability p_m. The value of the selected genes are changed by valid values depending of its locus, e.g. $g_1^j \in \{0, 1, 2\}$. Finally, the offspring was



parents

uniform crossover

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
2	5	32	0	0.5	8	3	32	0

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
2	**10**	**8**	0	0.5	**16**	3	32	**1**

→

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
1	10	8	1	0.2	16	5	64	1

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
1	**5**	**32**	1	0.2	**8**	5	64	**0**

flat and uniform Blx-alpha crossover

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
2	**10**	**8**	0	0.5	**16**	3	32	**1**

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
2	**10**	**16**	**1**	**0.4**	**16**	3	32	**1**

→

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
1	**5**	**32**	1	0.2	**8**	5	64	**0**

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
1	**5**	**32**	**1**	**0.3**	**8**	5	64	**0**

mutator

offsprings

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
2	**10**	**16**	**1**	**0.4**	**16**	3	32	**1**

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
1	**10**	**16**	**1**	**0.4**	**16**	3	**64**	**1**

→

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
1	**5**	**32**	**1**	**0.3**	**8**	5	64	**0**

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
0	**5**	**32**	**1**	**0.3**	**8**	5	**128**	**0**

Fig. 3. Example of the genetic operators applied in the genetic algorithm. Changed genes were highlighted in bold typeface.

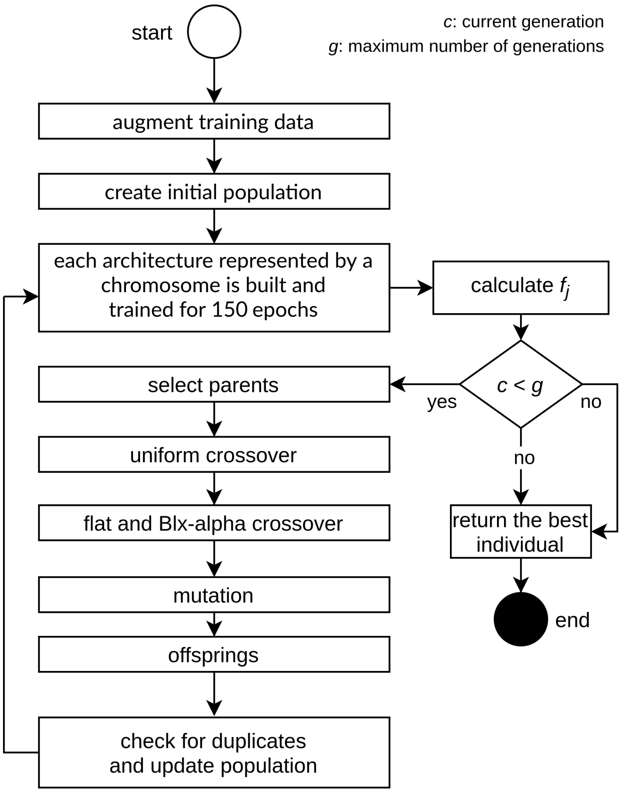

start
c: current generation
g: maximum number of generations
augment training data
create initial population
each architecture represented by a chromosome is built and trained for 150 epochs
calculate f_j
select parents
yes
c < g
no
uniform crossover
no
flat and Blx-alpha crossover
return the best individual
mutation
end
offsprings
check for duplicates and update population

Fig. 4. Integration of the GA and the training phase.

generated.

F. Population update

In order to update the population, a generational elitism approach was used [16]. The best parent replaces the worst child if the first one is better than all the children. After replacing the worst child, the new population replaces the previous one. As a consequence, the last generation will hold the best individual of the evolution.

G. Integration of the genetic algorithm in the training phase

Fig. 4 illustrates how the GA is integrated within the training phase of the model. Given an individual I_j, an architecture is built dynamically based on its chromosome. First, the gene g_1^j refers to which type of architecture will be build (see Fig. 1). Second, $g_p^j, p \in [2;6]$, represent the Feed Forward settings in Fig. 1b and Fig. 1c. Third, g_7^j and g_8^j refer to the Prediction block in Fig. 1a and Fig. 1b. Four, g_9^j means the applied optimizer - *Adaptive Moment Estimation* (ADAM) and *Stochastic Gradient Descent* (SGD) are used to solve the optimization problem, which iteratively minimizes the prediction errors on the training samples. This type of optimization algorithm traverses a path from a given initial point to near the optimum. Five, data is randomly augmented by following the settings detailed in Table II. Six, the initial population is created and trained for e epochs. Seven, the fitness of each individual is calculated. Eight, parents selection, crossover, and mutation are performed. Nine, duplicates are removed and the population is updated with the offspring. Then, steps from seven to nine are repeated until the maximum number of generations is reached. Finally, the best individual is returned.

TABLE I
SUMMARY OF THE ISIC2019 DATASET IN EACH FOLD. TRAIN DATA WAS AUGMENTED, AND TEST DATA KEPT THE ORIGINAL RATIO BETWEEN THE CATEGORIES.

Data	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	Total
Train	9,585	9,108	8,388	9,540	9,408	7,814	8,760	9,503	72,106
Test	1,045	86	236	23	26	437	325	63	2,241

TABLE II
CONFIGURATION USED DURING THE EXPERIMENTAL STUDY.

Parameter	Value
Population (p)	50
Uniform crossover probability (p_{uc})	90%
Flat crossover probability (p_{fc})	20%
Mutation probability (p_m)	20%
Number of generations (g)	50
Rotations	[1°,270°]
Flip	X-axis and Y-axis
Crop	[10%,30%]
Number of epochs (e)	150
Mini-batch size	16
Learning rate (α)	0.001

III. EXPERIMENTAL STUDY

A. Dataset

Table I shows a summary of the data used in the experimental study. The data was obtained from *The International Skin Imaging Collaboration*[1] (ISIC) repository. ISIC2019 comprises images that belong to eight categories. First, the data was filtered - the 2,556 instances with missing meta-data attributes were removed. Second, after each fold partition, training data were balanced by creating new images until the ratio of images was equal to the biggest one. The new training images were considered as independent from the original ones. On the other hand, test data was not augmented in order to keep the original ratio. Third, meta-data was composed by sex, age, and anatomical site. Finally, a total of ≈ 72,106 and 2,241 raw images were used in each fold for training and testing, respectively.

B. Deep convolutional neural network model

MobileNet was selected to analyze the images. This architecture uses depthwise separable convolutions, which significantly reduces the model size and complexity for training. The model requires images with a resolution of h=224, w=224, and c=3; h is the height, w is the width, and c is the number of channels.

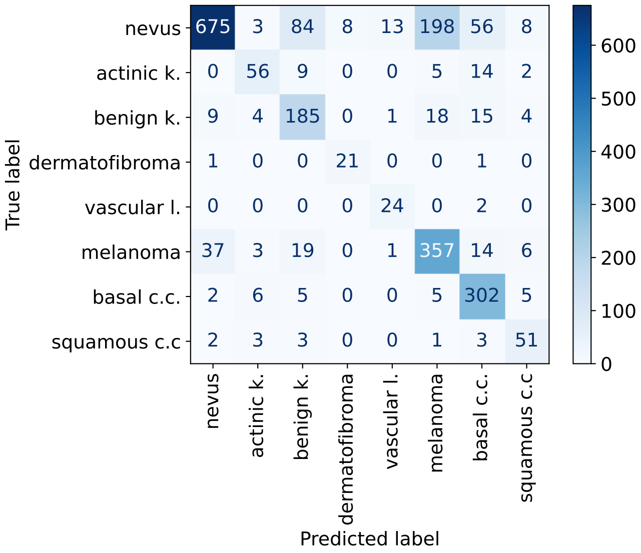


Fig. 5. Confusion matrix from the baseline MobileNet.

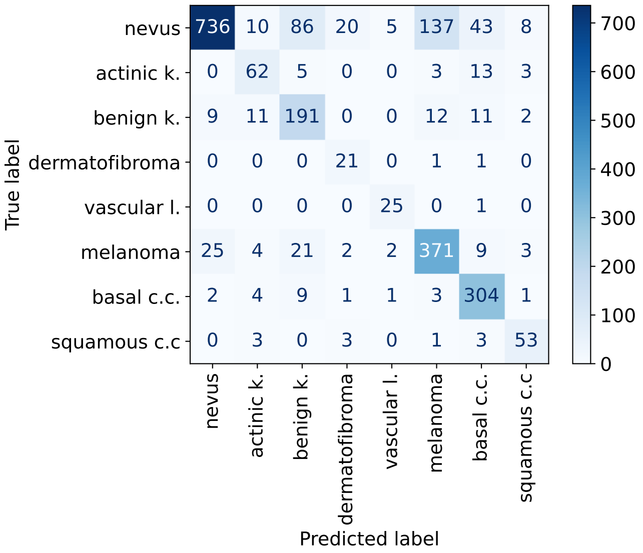


Fig. 6. Confusion matrix from the discovered architecture.

C. Experimental settings

Table II summarizes the basic configuration used throughout the experimental study to train the proposal. Data augmentation was performed as follows: random rotation angles between between 1° and 270°, random flips in the X-axis or Y-axis, and random cropping between 10% and 30% were used. Regarding the genetic algorithm that learn the best sets of architectures, the population was composed by 50 individuals and they were randomly initialized. The crossovers and mutator operators were applied with 90%, 20%, and 20% of probabilities, respectively. The individuals were evolved during 50 generations. As for the parameters used for training the deep learning models, the learning rate (α) was equal to 0.001 and it was reduced by a factor of 0.1 if an improvement in predictive performance was not observed during 10 epochs. Transfer learnig was not used in order to focus only in data augmentation techniques. The weights of the networks were initialized using Kaiming method [17] when not using transfer learning; a batch size equal to 16 was used; and the models were trained along 150 epochs. Finally, the cost function used for training the models was defined as the average of the categorical cross entropy along all training samples.

D. Evaluation process

Regarding the evaluation process, a 3-times 10-fold cross validation process was performed to assess the effectiveness of each model, and the results were averaged across all fold executions. In each fold, MCC was also used to measure the predictive performance of the models; the higher the MCC value, the better the performance. A perfect score is represented by 1.

TABLE III
RECALL SCORE OF EACH CLASS, AND FINAL MCC VALUE. THE BEST VALUE FOR EACH CLASS WAS HIGHLIGHTED IN BOLD TYPEFACE.

Model	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	MCC
Baseline	0.65	0.65	0.78	**0.91**	0.92	0.82	0.93	0.81	0.68
Ours	**0.70**	**0.72**	**0.81**	**0.91**	**0.96**	**0.85**	**0.94**	**0.84**	**0.73**

E. Software and hardware

Pytorch[2] was used as framework to build the proposed multi-view deep neural network architectures. DEAP [18] was used to develop the GA. The experimental study was executed in 4 × GPUs NVIDIA Geforce RTX 2080-Ti, 4 × GPUs NVIDIA Geforce RTX 1080-Ti, 2 × GPUs NVIDIA Geforce GTX 780, Ubuntu 18.04 and 20.04, and CUDA v11.

F. Results

First, the architectures discovered by the genetic algorithm were analyzed. The Borda's method was used to find which type of architecture achieved better overall performance in the entire GA. Borda is the simplest ranking aggregation method that assigns a score to an element in correspondence to the position in which this element appears in each ranking. Late fusion ended first (ranking 1875), followed by joint fusion (ranking 1600), and lastly early fusion (ranking 738). Nevertheless, a joint fusion architecture was the best of all the population, which is encoded as [1,5,8,1,0.6,16,1,128,0], and following results belong to it. Regarding the running time, we took the above settings and evaluated it with the three architectures in order to get an approximate running cost. Table IV shows how many seconds it takes to train with 72,106 images. As can be seen, the inclusion of a feed forward network takes its cost. Nevertheless, it is not a significant amount of time.

[1] https://www.isic-archive.com

[2] https://pytorch.org/

TABLE IV
RUNNING TIME FOR EACH EPOCH (SECONDS).

Model	time
Early fusion	**271**
Joint fusion	276
Late fusion	277

The confusion matrix for the baseline MobileNet is shown in Fig. 5, and for the selected discovered architecture in Fig. 6. Although the proposed GA is limited by the number of generations, the results show that better performance was attained when applying the GA than when using the baseline model. In addition, the last three categories indicate malignant skin lesions (melanoma, basal c.c. and squamous c.c.), where the proposed model was able to identify 18 more cases than the baseline. On the other, the predictive capability of the rest of categories was also improved, with a total of 74 new accurately identified lesions. Both models achieved the same performance when diagnosing dermatofibroma. In total, 92 more lesions are correctly classified by using the proposed architecture. It should be insightful to apply the proposal on non-dermoscopic images. It is well-known that the amount of this type of image is increasing sharply, leading to the possibility of embedding models on mobile devices. As a result, a preventive and more economical medical service is very likely to be boosted. Overall, nevus lesions are still the more complex images, given their great variety and similarity with the rest of the categories.

Table III summarizes the performance of the models in each category. Vascular lesion, basal c.c. and dermatofibroma were the three most clearly identifiable categories, with 96%, 94%, and 91% of recall, respectively. The proposal overcame the baseline in all categories, except in dermatofibroma, and it achieved the best MCC score, surpassing the baseline by 5 points. The results corroborated that meta-data is more useful after been processed by a feed forward network, which can be considered as a key finding. Abstract features extracted by the network poses as the main difference between early feature-level fusion and the other two approaches that used two training models. Also, the joint fusion architecture that was trained by using the loss from the final prediction, achieved the top performance. The above point out to a regularization effect between both models.

IV. Conclusion and future work

In this work, a novel multimodal fusion architecture for the automatic diagnosis of multiclass skin lesions is discovered by using a genetic algorithm. This work clearly states the aim of proving or not the validity of using meta-data, which has been corroborated by the experimental study as a suitable approach. As a result, the foundation to explore different ways for finding a better performance has been laid, or even in a shorter computation time. The proposal follows a multi-view-network approach, where each model learns from training images, clinical data, or fused abstract features extracted from the mentioned sources. The proposal was competitive with the baseline state-of-the-art model. All categories were improved, except for a tie while diagnosing dermatofibroma. Future works will include more state-of-the-art CNN models and datasets in order to validate that the genetic algorithm is still able to achieve suitable performance. Also, the impact of the GA's hyperparameters on the model's performance should be analyzed, which requires a very expensive work. Finally, it is worth noting that the design of custom crossover methods could lead to improve the performance, e.g. by considering blocks of genes to evolve instead of single genes.

References

[1] American Cancer Society, "Cancer Facts and Figures," 2022, consulted on February 6, 2022. [Online]. Available: https://bit.ly/3HvQhuQ

[2] H. D. Lee, A. I. Mendes, N. Spolaôr, J. T. Oliva, A. R. Sabino Parmezan, F. C. Wu, and R. Fonseca-Pinto, "Dermoscopic assisted diagnosis in melanoma: Reviewing results, optimizing methodologies and quantifying empirical guidelines," *Knowledge-Based Systems*, vol. 158, pp. 9–24, oct 2018.

[3] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[4] W. Abbes and D. Sellami, "High-level features for automatic skin lesions neural network based classification," in *Proceedings of the 2nd International Image Processing, Applications and Systems Conference*, Hammamet, Tunisia, 2017.

[5] H. Haenssle et al., "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[6] E. Pérez and S. Ventura, "Melanoma Recognition by Fusing Convolutional Blocks and Dynamic Routing between Capsules," *Cancers*, vol. 13, no. 19, 2021.

[7] X. Zhang et al., "Cmc: A consensus multi-view clustering model for predicting alzheimer's disease progression," *Computer Methods and Programs in Biomedicine*, vol. 199, 2021.

[8] D. Ramachandram and G. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[9] J. Kawahara et al., "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, 2019.

[10] E. Pérez and S. Ventura, "An ensemble-based convolutional neural network model powered by a genetic algorithm for melanoma diagnosis," *Neural Computing and Applications*, 2021.

[11] C. Yifei et al., "Computer aided diagnosis of thyroid nodules based on the devised small-datasets multi-view ensemble learning," *Medical Image Analysis*, vol. 67, p. 101819, 2021.

[12] E. Pérez, O. Reyes, and S. Ventura, "Convolutional neural networks for the automatic diagnosis of melanoma: An extensive experimental study," *Medical Image Analysis*, vol. 67, 2021.

[13] D. Chicco, N. Tötsch, and G. Jurman, "The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Mining*, vol. 14, pp. 1–22, 2021.

[14] K. Deb, "Genetic algorithms for function optimisation," *Genetic algorithms and soft computing*, vol. 8, pp. 4–31, 1996.

[15] F. Herrera, M. Lozano, and J. L. Verdegay, "Tackling Real-Coded Genetic Algorithms: Operators and Tools for Behavioural Analysis," *Artif Intell Rev*, vol. 12, no. 4, pp. 265–319, 1998.

[16] T. Bäck, *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms.* USA: Oxford University Press, Inc., 1996.

[17] K. He et al., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[18] F. Félix-Antoine et al., "DEAP: Evolutionary algorithms made easy," *Journal of Machine Learning Research*, vol. 13, pp. 2171–2175, jul 2012.

7.2. Data augmentation via a multi-view architecture and a genetic algorithm

Title	Performing melanoma diagnosis by an effective multi-view convolutional architecture
Authors	O. Reyes, E. Pérez, S. Ventura
Contest	The International Skin Imaging Collaboration challenge (ISIC): Skin Lesion Analysis Towards Melanoma Detection. 2019.
Year	2019
Status	Accepted

PERFORMING MELANOMA DIAGNOSIS BY AN EFFECTIVE MULTI-VIEW CONVOLUTIONAL ARCHITECTURE

A PREPRINT

Oscar Reyes, Eduardo Pérez, Sebastián Ventura*
Department of Computer Science and Numerical Analysis, University of Córdoba.
Knowledge Discovery and Intelligent Systems in Biomedicine Laboratory
Maimónides Biomedical Research Institute of Córdoba.
Córdoba, Spain
ogreyes@uco.es, eduardo.perez@imibic.org, sventura@uco.es

August 15, 2019

1 Introduction

Melanoma is a major public health problem since it has the highest levels of mortality between the different types of skin cancer [1]. Alarmingly, this illness has an increasing incidence in white people, where just in Europe were estimated 144,200 cases and 20,000 deaths in 2018 [2]. Consequently, global actions are needed to revitalize efforts for melanoma control and prevention.

Despite the progress achieved in clinical tests, the early diagnosis of melanoma remains as a tough task even for expert dermatologists because the complexity, variability and dubiousness of the symptoms [3]. In this regard, several studies have shown that the early diagnosis of melanoma can be benefited from computational methods [4], and recent works demonstrated that such techniques may even overcome the diagnosis made by a committee of expert dermatologists [5]. Classical machine learning techniques commonly require the previous extraction of handcrafted features, thus incorporating a priori an important knowledge in the analysis, and not requiring the availability of very large datasets for constructing a proper predictive model [6]. However, despite the advantages of these techniques, it should be noted that the quality of the extracted descriptors heavily relies on the level of dermatologists' expertise, and finding the informative and discriminative set of high-level features to build an accurate model remains as a complex and costly task that is usually problem dependent [7]. Consequently, in the last years there has been an increasing attention in developing computational techniques which can automatically extract and learn high-level features, thus providing a higher robustness to the inter- and intra-class variability present in melanoma images [8].

Deep learning models, specifically Convolutional Neural Networks (CNNs), are widely being been used for melanoma diagnosis from dermoscopic images [9]. This type of learning method has the capacity of automatically learning high-level features from raw images [8], thus allowing the extraction of hierarchies of features by applying convolutional operators that progressively learn more abstract features and, finally, enabling the learning of data-driven features for specific tasks [10]. Therefore, CNNs has shown to be more effective in diagnosing melanoma, easing the development of novel applications in a shorter time.

Accordingly, this work presents a novel deep learning model for diagnosing melanoma from images. The proposed architecture is inspired in multi-view paradigm, where several views were generated from the original images, allowing to achieve a promising effectiveness in melanoma diagnosis. Next, the proposed model is briefly explained, and the specific configuration used in *The International Skin Imaging Collaboration* 2019 challenge (ISIC-2019) is also portrayed.

[*] This work is copyrighted by Dr. O. Reyes, MSc. E. Pérez, and Dr. S. Ventura, and any copy, modification, and posterior publication by other authors is not allowed. This manuscript was submitted to ISIC-2019 with the only intention to briefly describe the model used in the competition.

2 Description of the proposed approach

In this work, a novel architecture inspired in multi-view learning for melanoma diagnosis was proposed. Figure 1 briefly describes the proposed architecture, which is based on the following hypothesis: *better abstract and discriminative features would be extracted if independent computational blocks were learned from restricted and specialized feature spaces*. Each different view of an image is constructed by means of applying independent and specialized random transformations over the same original image. Therefore, each view represents an unique feature space, thus reducing the large number of representations that would be generated if all possible random transformations were simultaneously applied. A view is generated by applying a group of random transformations belonging to the same category (e.g. colour-based transformations) and, therefore, we can produce a desired number of different and independent views from the same image.

The architecture is composed by N independent computational blocks which learn from the different views generated for an image, where each block yields a partial prediction (p_i) and its learned representation. Also, the new representations produced by each block are then concatenated by channels and passed to a final block that produces an additional prediction. Finally, given a sample image, the model predicts its class by using an aggregation procedure that considers the $N + 1$ partial predictions ($\{p_1, p_2, \dots, p_N, p_{N+1}\}$).

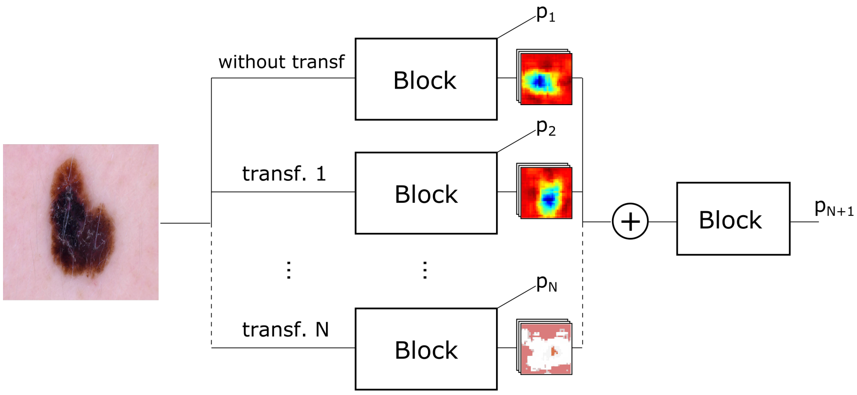


Figure 1: The proposed multi-view architecture.

The proposed architecture have several advantages, to mention a few:

- The model implicitly follows an ensemble approach which can benefit the predictive performance in melanoma diagnosis. Ensemble-based learning has demonstrated to be an effective way to improve the learning process in many real-world problems [11], and melanoma diagnosis is no exception [12].
- The model uses the so-called data augmentation technique for constructing the different views of an image. Therefore, data are implicitly augmented not only at training phase, but also at test phase, allowing to attain a better predictive performance; the application of data augmentation on both phases have shown to be an effective way to improve melanoma diagnosis [13]. On the other hand, data augmentation technique has demonstrated to be an effective regularization method [14].
- Each computational block can extract richer abstract features which are focused on specific feature spaces; each view is generated after applying random operations which are grouped by types of transformations. This feature can also reduce the complexity in the learning process of each computational block, and therefore the overall computational cost of the model, since each block is learning from simpler feature spaces.
- The model not only learns specific abstract features that describes each independent view, but it can also be benefited from those shared features that exist across the views; the final block can extract features from the representation that was formed by combining the learned feature spaces from each intermediate block.
- The model produces the final prediction by considering several auxiliary classifiers. This feature can yield a more stable learning and better convergence [15], as well as the auxiliary classifiers can act as regularizer [16].
- Finally, the architecture is flexible with respect to the designing of each internal block. We can use from single stacked convolutional layers to a complex independent model (e.g. an Inception model) in each block. Also, blocks comprising the same internal structure can be used along the architecture, or we can use a collection of heterogeneous blocks. This flexibility gives the chance in defining a diverse ensemble of independent models, having each one different advantages and drawbacks.

3 Specific configuration used in ISIC-2019 challenge

This was the configuration used of the proposed architecture in ISIC-2019:

- The model had five intermediate blocks, where the first one used the original images (i.e. no transformations were applied), and the rest of blocks learned from different views, being each one generated by a different group of random transformations. The four family of random transformations were: rotation-based transformations, flip-based transformations, crop-based transformations, and colour-based transformations.
- A pretained MobileNetV2 [17] was used in each intermediate block. This model had weights that were previosly learned from ImageNet [18] dataset, which contains more than 1 million of images of 1,000 object classes. Also, given an image, MobileNetV2 was modified in order to not only obtain the prediction for this image, but also the learned representation that is subsequently concatenated and passed to the final block of the architecture.
- The last block of the architecture comprises three sub-blocks, where each one encompasses the following stacked layers: **convolutional** -> **batch-normalization** -> **Relu** -> **dropout** (probability of 0.4). Finally, there is an additional convolucional layer followed by a flatten layer that gives the prediction.
- A cost function that averages a softmax cross entropy (loss function) across all the blocks was used to train the network.
- The final predictions were obtained combining all the partial predictions yielded by the blocks using a soft-voting approach.
- The model was trained with the optimizer *Stocastic Gradient Descend* (SGD) using a batch size equal to 16. The default parameters proposed for SGD algorithm were used.
- The model was trained along 50 epochs.
- Considering that the intermediate blocks are composed by pretained MobileNetV2, in this work we applied the following fine tuning approach: all the internal layers of each MobileNetV2, excepting those last layers that were modified to directly obtain the learned representations, were “frozen" during the first five epochs. After five epochs, all the blocks, including all their internal layers, were defrosted.
- To measure the effectiveness of the proposed model, as well as for tuning the main hyperparameters, a 10-fold cross validation was conducted on the training set.
- The model was executed in a Desktop PC with Ubuntu 18.04, Intel Core i7-8700K Processor 3.7 GHz, 64 GB DDR4 RAM, and two GPUs Geforce GTX 1080-Ti with 11 GB DDR5 each one. We used Gluon library of Apache MXNet[2], which provides a clear, simple and concise API for designing deep learning models.

For the second task of the challenge, *Lession Diagnosis: Images and Metadata*, the same approach used in the task no. 1 was used, but in this case an additional view was generated from the provided meta-data. The predictions yielded by the view related with meta-data is integrated into the model by averaging them with the ones given by the other image-based views. In order to meta-data could directly be used by the neural network, the categorical variables were converted to dummy variables, the variable *age* was normalized into $[0, 1]$ range, and all the missing values were replaced by the median or mode of the corresponding variable. The intermediate block that learns from meta-data was a simple multi-layer perceptron composed by three dense hidden layers of 32 ReLU units each one, and a final output layer with a number of units equal to the number of existing classes; a dropout layer was placed after each hidden layer using a 0.4 probability.

Finally, in the two tasks, the following two approaches were used to determine when the category of an image should be considered as unknown; all these approaches are based on active learning paradigm:

- Entropy sampling approach: given an image, the uncertainty in the prediction yielded by the model was measured by means of computing its entropy. The more uniform is the distribution of probabilities, the higher the uncertainty in the prediction. Those samples with a normalized entropy greater than 0.85 were directly considered as unknown.
- Margin sampling: given an image, the uncertainty in the prediction yielded by the model was measured by means of computing the difference between the two more likely classes. The smaller the difference between the two highest probabilities, the highest the uncertainty in the prediction. Those samples with a margin distance lower than 0.02 were directly considered as unknown.

[2] https://mxnet.incubator.apache.org/

- Relevance sampling: given an image, the uncertainty in the prediction yielded by the model was measured by means of computing the representativeness of the more likely class. The smaller the probability of the more likely class, the highest the uncertainty in the prediction. Those samples with a more likely class not comprising at least 30% of the total of the distribution were directly considered as unknown.

Acknowledgments

This research was supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund, projects TIN2017-83445-P. It was also supported by the i-PFIS contract no. IFI17/00015 granted by Health Institute Carlos III of Spain.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2019,” *CA Cancer Journal for Clinicians*, vol. 69, pp. 7–34, 2019.

[2] J. Ferlay, M. Colombet, I. Soerjomataram, T. Dyba, G. Randi, M. Bettio, A. Gavin, O. Visser, and F. Bray, “Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018,” *European Journal of Cancer*, vol. 103, pp. 356–387, 2018.

[3] A. C. Geller, S. M. Swetter, K. Brooks, M.-F. Demierre, and A. L. Yaroch, “Screening, early detection, and trends for melanoma: Current status (2000-2006) and future directions,” *Journal of the American Academy of Dermatology*, vol. 57, no. 4, pp. 555–572, 2007.

[4] H. D. Lee, A. I. Mendes, N. Spolaôr, J. T. Oliva, A. R. Sabino Parmezan, F. C. Wu, and R. Fonseca-Pinto, “Dermoscopic assisted diagnosis in melanoma: Reviewing results, optimizing methodologies and quantifying empirical guidelines,” *Knowledge-Based Systems*, vol. 158, pp. 9–24, oct 2018.

[5] H. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, and L. Uhlmann, “Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists,” *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[6] J. Sánchez-Monedero, M. Pérez-Ortiz, A. Sáez, P. A. Gutiérrez, and C. Hervás-Martínez, “Partial order label decomposition approaches for melanoma diagnosis,” *Applied Soft Computing Journal*, vol. 64, pp. 341–355, 2018.

[7] L. Jin, S. Gao, Z. Li, and J. Tang, “Hand-crafted features or machine learnt features? together they improve RGB-D object recognition,” in *Proceedings - 2014 IEEE International Symposium on Multimedia, ISM 2014*, 2015, pp. 311–319.

[8] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[9] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun, “Deep learning for image-based cancer detection and diagnosis A survey,” *Pattern Recognition*, vol. 83, pp. 134–149, 2018.

[10] X. Zhen, L. Shao, S. J. Maybank, and R. Chellappa, “Handcrafted vs. learned representations for human action recognition,” *Image and Vision Computing*, vol. 55, pp. 39–41, 2016.

[11] L. Rokach, “Ensemble-based classifiers,” *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.

[12] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga, “Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble,” *arXiv preprint arXiv:1703.03108*, 2017.

[13] A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. Avila, and E. Valle, “RECOD Titans at ISIC Challenge 2017,” 2017.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[15] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, “Deeply-supervised nets,” in *Artificial intelligence and statistics*, 2015, pp. 562–570.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.

[18] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, vol. 2, 2012, pp. 1097–1105.

7.3. Investigación y recopilación preliminar de los métodos más prometedores para diagnosticar melanoma

UNIVERSIDAD
Đ
CÓRDOBA

Title	Diagnóstico automático de melanoma: una revisión
Authors	E. Pérez, O. Reyes, S. Ventura
Conference	VII Congreso Científico de Investigadores en Formación
Year	2019
Editorial	UCOPress, Editorial Universidad de Córdoba
Status	Accepted

Diagnóstico automático de melanoma: una revisión

Eduardo Pérez Perdomo, Oscar Gabriel Reyes Pupo, Sebastián Ventura Soto

Universidad de Córdoba. Escuela Politécnica Superior de Córdoba. Departamento de Informática y Análisis Numérico. E-mail: eduardo.perez@imibic.org, ogreyes@uco.es, sventura@uco.es

Summary

Melanoma is the type of skin cancer with a major probability to propagate to nearby tissues, and it is today the skin cancer which has the highest levels of mortality. The melanoma diagnosis is a complex task, even for an expert dermatologist, due to the high variability in the characteristics of moles that appear in the patients' skin. Several studies have demonstrated that machine learning techniques can be really helpful and effective for automatic melanoma diagnosis. In this work, we aimed to study the main machine learning methods that have been proposed so far for diagnosing melanoma from dermoscopic images. The bibliography revision revealed the most effective state-of-the-art techniques to perform the melanoma diagnosis, and it also allowed to detect the main challenges that still persist in diagnosing melanoma from images.

Resumen

El melanoma es el tipo de cáncer de piel que tiene mayor probabilidad de invadir el tejido cercano y diseminarse a otras partes del cuerpo, siendo hoy en día el cáncer de piel que provoca más muertes. El diagnóstico de melanoma es una tarea compleja incluso para un dermatólogo experto debido a la alta variabilidad en las características de los lunares que aparecen en las pieles de los pacientes. Varios trabajos han demostrado que las técnicas de aprendizaje automático pueden ser realmente útiles para el diagnóstico automático de melanoma. El objetivo principal de este trabajo fue realizar un estudio bibliográfico de las principales técnicas de aprendizaje automático que han sido propuestas para el diagnóstico de melanoma a partir de imágenes dermoscópicas. A partir de la revisión bibliográfica se determinaron cuáles son las técnicas más efectivas, así como los principales desafíos que aún subsisten para un correcto diagnóstico automático de melanoma.

Introducción

Actualmente se diagnostican alrededor de todo el mundo unos 160,000 casos al año que padecen cáncer de piel [1], y de ese total solo en España se diagnostican unos 3,600 casos. Varios estudios han demostrado que la mayoría de las consecuencias graves que produce el melanoma se pueden evitar mediante su diagnóstico temprano [2]. Inicialmente, para determinar si un paciente padece de melanoma o no, se estudian las imágenes dermoscópicas de las lesiones, analizando el color, la geometría y la textura del área dañada [3]. Sin embargo, el diagnóstico de melanoma es una tarea compleja incluso para un dermatólogo experto debido a la alta variabilidad en las características de las imágenes; como se puede observar en las Figuras 1 y 2 existen una diferencia significativa entre imágenes pertenecientes a un mismo grupo.

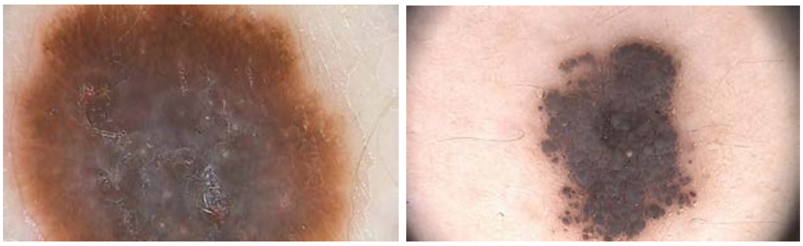


Figura 1. Muestras sanas.

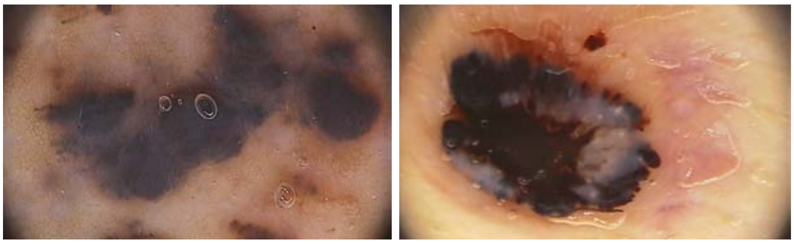


Figura 2. Melanoma.

A lo largo de los años, varios trabajos han demostrado que los algoritmos de aprendizaje automático pueden favorecer significativamente el proceso de diagnóstico de melanoma a partir de imágenes [4]. El objetivo de este trabajo fue realizar un estudio bibliográfico que tiene como objeto de estudio el *diagnóstico de melanoma*, enfocándose en el campo de acción correspondiente al *diagnóstico mediante técnicas de aprendizaje automático*. A partir del estudio bibliográfico realizado, se determinaron las técnicas computacionales más usadas en la actualidad para el diagnóstico automático de melanoma, las cuales servirán posteriormente como punto comparativo para validar los métodos que nos planteamos desarrollar a lo largo de la tesis doctoral.

Metodología

Para desarrollar este trabajo fueron consultadas las principales publicaciones, en revistas especializadas, relacionadas con el diagnóstico automático de melanoma en los últimos 13 años, así como las principales conferencias internacionales relacionadas con la temática. Para la búsqueda de la bibliografía especializada utilizamos principalmente la base de datos SCOPUS, empleando diferentes combinaciones de palabras claves para la búsqueda, tales como AUTOMATIC SKIN CANCER DETECTION/DIAGNOSIS, AUTOMATED MELANOMA DETECTION/DIAGNOSIS, MACHINE LEARNING & MELANOMA, DEEP LEARNING & MELANOMA, etc. Esta búsqueda nos permitió filtrar los artículos relacionados por citas y año de publicación, detectando el estado del arte más reciente en la temática, y dejando de esta manera sentadas las bases para la continuidad de la investigación.

Resultados

A partir del estudio bibliográfico realizado pudimos observar que la mayoría de los métodos computacionales existentes utilizan como entrada las imágenes de los lunares de cada paciente, o conjunto de descriptores extraídos de esas imágenes, y aplican técnicas clásicas como los *k* vecinos más cercanos [5], redes neuronales artificiales [6] y máquinas de soporte vectoriales [7]. La principal limitación de estas técnicas radica en que dependen de la calidad de los descriptores que se extraen de las imágenes, o que las imágenes originales se convierten en vectores de gran longitud; esta última limitación trae consigo el desafío de construir un modelo preciso a partir de pocas muestras (imágenes) que son descritas por un enorme número de variables.

Por otra parte, en los últimos años se ha observado un especial interés en la aplicación de modelos de redes profundas [8], los cuales han demostrado ser realmente efectivos en la clasificación de imágenes en problemas de diferentes dominios de aplicación. La efectividad de los modelos profundos para la detección de melanoma se puede evidenciar en las competiciones que se viene desarrollando desde el 2016 patrocinadas por la fundación *The International Skin Imaging Collaboration* (ISIC), donde las *Convolutional Neural Networks* (CNN) han ocupado los primeros lugares. Las CNN tienen como ventaja que aprenden directamente a partir de las imágenes originales, por lo que no se ven afectadas por transformaciones realizadas en etapas de preprocesamiento. Estos modelos son capaces de extraer patrones espaciales abstractos a lo largo de todas las capas ocultas de la red, los cuales son usados posteriormente para clasificar las imágenes. Sin embargo, aún quedan varios desafíos en el área del diagnóstico del melanoma a partir de imágenes, como por ejemplo lograr construir modelos precisos donde el desbalance entre el número de muestras de cada clase es elevado y/o el tamaño de las bases de datos de imágenes de melanoma es pequeño. Ejemplo de ello son los resultados obtenidos en ISIC-2016 e ISIC-2017, mostrados en la Tabla 1, los cuales evidencian que los niveles de predicción alcanzados en la clase positiva son aún bajos, confirmando que los modelos no son robustos ante el desbalance existente entre el número de muestras de cada clase.

Tabla 1. Resultados obtenidos por los ganadores de ISIC en la tarea de clasificación de melanoma.

	Exactitud	TPR	TNR	Desbalance	ISIC
CUMED [9]	85,50	50,70	94,10	4,1x	2016
Resnet ensemble [10]	82,80	73,50	85,10	4,3x	2017

Conclusiones

A partir del estudio bibliográfico realizado se identificaron las principales técnicas de aprendizaje automático empleadas para el diagnóstico automático de melanoma. Se comprobó que la eficacia de las técnicas que incluyen una etapa de preprocesamiento de las imágenes dermoscópicas depende en gran medida de la calidad de los descriptores extraídos a partir de las imágenes. Por otro lado, se observó que en los últimos años las redes profundos se han convertido en los modelos más utilizados para el diagnóstico de melanoma, en parte por su capacidad de trabajar directamente con las imágenes originales. Sin embargo, a pesar de la alta capacidad de los modelos profundos, se detectaron una serie de desafíos que aún persisten en la tarea del diagnóstico de melanoma. En trabajos futuros nos proponemos realizar estudios experimentales para el análisis de la efectividad de las diferentes estructuras de redes convolucionales. También diseñaremos estrategias para disminuir el impacto negativo que produce el desbalance en los datos, así como abordaremos el problema del entrenamiento de redes profundas a partir de base de datos de tamaño pequeño.

Bibliografía

[1] Evolución del melanoma. Retrieved November 29, 2018, from https://www.aecc.es/es/todo-sobre-cancer/tipos-cancer/melanoma/evolucion-melanoma.

[2] Rogers, H.W., Weinstock, M.A., Feldman, S.R. and Coldiron, B.M.. Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population, 2012. JAMA dermatology , 151 (10), 2015, p. 1081-1086.

[3] Sánchez-Monedero, J., Pérez-Ortiz, M., Sáez, A., Gutiérrez, P.A. and Hervás-Martínez, C.. Partial order label decomposition approaches for melanoma diagnosis. Appl Soft Comput, 64, 2018, p. 341-355.

[4] M. A. Arasi, E. A. El-Dahshan, E. M. El-Horbaty and A. M. Salem. Malignant Melanoma Detection Based on Machine Learning Techniques: A Survey. Egyptian Computer Science Journal, 40 (3), 2016.

[5] L. Li, Q. Zhang, Y. Ding, H. Jiang, Bruce H. Thiers and J. Z. Wang. Automatic diagnosis of melanoma using machine learning methods on a spectroscopic system. BMC Med Imaging, 2014, p. 14-36.

[6] M. Barzegari, H. Ghaninezhad, P. Mansoori, A. Taheri, Z. S. Naraghi and M. Asgari. Computer-Aided dermoscopy for diagnosis of melanoma. BMC Dermatology, 5 (1), 2005.

[7] S. Gilmore, R. Hofmann-Wellenhof and H. Soyer. A support vector machine for decision support in melanoma recognition. Exp Dermatol, 19 (9), 2010, p. 830-835.

[8] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542 (7639), 2017, p. 115-118.

[9] Yu, L., Chen, H., Dou, Q., Qin, J. and Heng, P.A.. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE transactions on medical imaging, 36(4), 2017, p. 994-1004.

[10] Matsunaga, K., Hamada, A., Minagawa, A., and Koga, H. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. ArXiv Preprint ArXiv:1703.03108, 2017.

Agradecimientos

Este trabajo ha sido financiado por el contrato i-PFIS no. IFI17/00015 otorgado por el Instituto de Salud Carlos III. También ha sido financiado por el proyecto TIN2017-83445-P del Ministerio de Economía y Competitividad y Fondos FEDER.

DIAGNÓSTICO AUTOMÁTICO DE MELANOMA: UNA REVISIÓN

Eduardo Pérez Perdomo[1], Oscar Gabriel Reyes Pupo[2], Sebastián Ventura Soto[2]
[1] Instituto Maimónides de Investigación Biomédica de Córdoba
eduardo.perez@imibic.org
[2] Universidad de Córdoba. Escuela Politécnica Superior de Córdoba. Departamento de Informática y Análisis Numérico
{ogreyes, sventura}@uco.es

Introducción

El melanoma es el tipo de cáncer de piel que tiene mayor probabilidad de invadir el tejido cercano y diseminarse a otras partes del cuerpo, siendo hoy en día el cáncer de piel con mayor mortalidad. El diagnóstico de melanoma es una tarea compleja incluso para un dermatólogo experto, debido a la alta variabilidad en las características de los lunares que aparecen en la piel de los pacientes. Actualmente se diagnostican alrededor de todo el mundo unos 160.000 casos de cáncer de piel, y de ese total solo en España se diagnostican unos 3.600 casos. Para determinar si un paciente padece o no esta enfermedad, se estudian las imágenes dermoscópicas de las lesiones, analizando el color, la geometría y la textura del área dañada.

Resultados

La efectividad de los modelos profundos para la detección de melanoma se puede evidenciar en las competiciones que se vienen desarrollando desde el 2016 patrocinadas por la fundación The International Skin Imaging Collaboration (ISIC), donde las Convolutional Neural Networks (CNN) han ocupado los primeros lugares. Los modelos CNN son capaces de extraer patrones espaciales abstractos a lo largo de todas las capas ocultas de la red, los cuales son usados posteriormente para clasificar las imágenes. Los resultados obtenidos en ISIC-2016 e ISIC-2017 se muestran en la Tabla 1.

	Exactitud	TPR	TNR	Desbalance	ISIC
CUMED	85,50	50,70	94,10	4,1x	2016
Resnetensemble	82,80	73,50	85,10	4,3x	2017

Tabla 1. Resultados obtenidos por los ganadores de ISIC en la tarea de clasificación de la lesión.

Conclusiones

Se identificaron las principales técnicas de aprendizaje automático empleadas para el diagnóstico automático de melanoma. Se observó que en los últimos años las redes profundas se han convertido en los modelos más utilizados para el diagnóstico automático de melanoma, en parte por su capacidad de trabajar directamente con las imágenes originales. En trabajos futuros nos proponemos realizar estudios experimentales para el análisis de la efectividad de las diferentes estructuras de redes convolucionales. También diseñaremos estrategias para disminuir el impacto negativo que produce el desbalance en los datos y abordaremos el problema del entrenamiento de las redes profundas a partir de bases de datos de tamaño pequeño.

Agradecimientos

Este trabajo ha sido financiado por el contrato i-PFIS no. IFI17/00015 otorgado por el Instituto de Salud Carlos III. También ha sido financiado por el proyecto TIN2017-83445-P del Ministerio de Economía y Competitividad y Fondos FEDER.

Bibliografía

[1] Sánchez-Monedero, J., Pérez-Ortiz, M., Sáez, A., Gutiérrez, P.A. and Hervás-Martínez, C. Partial order label decomposition approaches for melanoma diagnosis. Appl Soft Comput, 64, 2018, p. 341-355.

[2] L. Li, Q. Zhang, Y. Ding, H. Jiang, Bruce H. Thiers and J. Z. Wang. Automatic diagnosis of melanoma using machine learning methods on a spectroscopic system. BMC Med Imaging, 2014, p. 14-36.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542 (7639), 2017, p. 115-118.

[4] Yu, L., Chen, H., Dou, Q., Qin, J. and Heng, P.A.. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE transactions on medical imaging, 36(4), 2017, p. 994-1004.

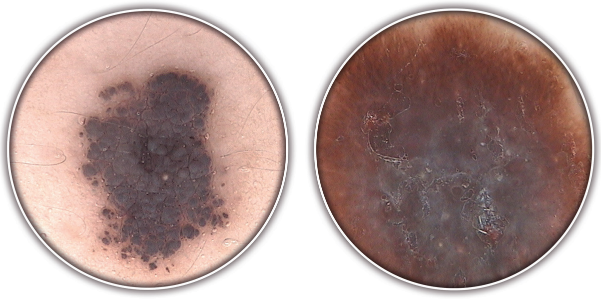


Figura 1. Muestras sanas.

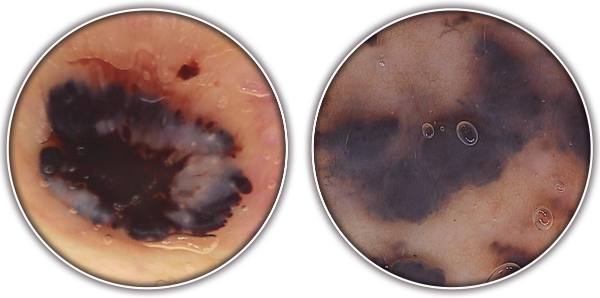


Figura 2. Melanomas

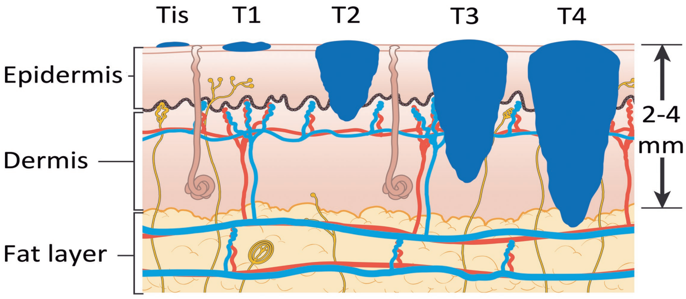
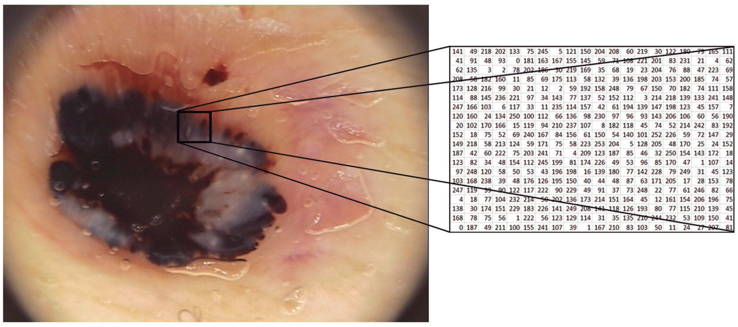


Figura 3. Etapas del Melanoma. Fuente: Cancer Research UK/Wikimedia Commons..



141	49	218	202	133	75	245	5	121	150	204	208	60	219	30	122	180	79	165	111
41	91	48	93	0	181	163	167	155	145	59	71	108	221	201	83	231	21	4	62
62	135	3	2	78	202	186	30	219	169	35	68	19	23	204	76	88	47	223	69
208	58	182	160	11	85	69	175	113	58	132	39	136	198	203	153	200	185	74	57
173	128	216	99	30	21	12	2	59	192	158	248	79	67	150	70	182	74	111	158
114	88	145	236	221	97	34	143	77	137	52	152	112	3	214	218	139	133	241	148
247	166	103	6	117	33	11	235	114	157	42	61	194	139	147	198	123	45	157	7
120	160	24	134	250	100	112	66	136	98	230	97	96	93	143	206	106	60	56	190
20	102	170	166	15	119	94	210	237	107	8	182	118	45	74	52	214	242	83	192
152	18	75	52	69	240	167	84	156	61	150	54	140	101	252	226	59	72	147	29
149	218	58	213	124	59	171	75	58	223	253	204	5	128	205	48	170	25	24	152
187	42	60	222	75	203	241	71	4	209	123	187	85	46	32	250	154	143	172	18
123	82	34	48	154	112	245	199	81	174	226	49	53	96	85	170	47	1	107	14
97	248	120	58	50	53	43	196	198	16	139	180	77	142	228	79	249	31	45	123
103	168	238	39	48	176	126	195	150	40	44	48	87	63	171	205	17	28	153	78
247	119	39	90	122	117	222	90	229	49	91	37	73	248	22	77	61	246	82	66
4	18	77	104	232	214	50	202	136	173	214	151	164	45	12	161	154	206	196	75
138	30	174	151	229	183	226	141	249	208	141	118	126	193	80	77	115	210	139	45
168	78	75	56	1	222	56	123	129	114	31	35	135	210	244	232	53	109	150	41
0	187	49	211	100	155	241	107	39	1	167	210	83	103	50	11	24	27	207	81

Figura 4. Lo que entiende el ordenador

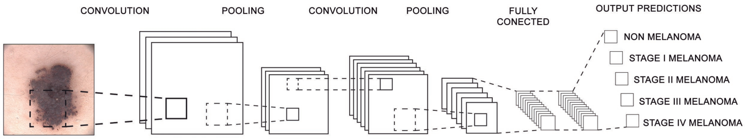


Figura 5. Estructura de una Convolutional Neural Networks (CNN).

Presentado en: VII Congreso Científico de Investigadores en Formación de la Universidad de Córdoba
Córdoba, 6 y 7 de febrero de 2019

8

Other conference publications

8.1. National conferences

1. E. Pérez, L. González, L. Sánchez, O. Reyes, S. Ventura. *JCLAL 2.0: mejoras y nuevas funcionalidades en la herramienta Java de código abierto para el aprendizaje activo.* Proceedings of the XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA), pp. 901-906. Granada (Spain). October 2018. Available from SCI2S digital library.

2. O. Reyes, J. Luna, J. Moyano, E. Pérez, S. Ventura. *Resolución de Problemas Biomédicos mediante Técnicas de Extracción de Conocimiento.* Proceedings of the XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA), pp. 1252-1257. Granada (Spain). October 2018. Available from SCI2S digital library.