ISSN 2442-6571

Systematic literature review of dermoscopic pigmented skin lesions classification using convolutional neural network (CNN)

Erwin Setyo Nugroho [a,b,1] , Igi Ardiyanto [a,2], Hanung Adi Nugroho [a,3,*]

[a] Department of Electrical and Information Engineering, Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia
[b] Department of Informatics, Politeknik Caltex, Riau, Indonesia
[1] erwinsn@pcr.ac.id; [2] igi@ugm.ac.id; [3] adinugroho@ugm.ac.id
* corresponding author

ARTICLE INFO

Article history
Received December 12, 2022
Revised April 10, 2023
Accepted May 21, 2023
Available online October 10, 2023

Keywords
Pigmented skin lesion
Dermoscopy
Melanoma
Skin cancer
Convolutional Neural Network

ABSTRACT

The occurrence of pigmented skin lesions (PSL), including melanoma, are rising, and early detection is crucial for reducing mortality. To assist Pigmented skin lesions, including melanoma, are rising, and early detection is crucial in reducing mortality. To aid dermatologists in early detection, computational techniques have been developed. This research conducted a systematic literature review (SLR) to identify research goals, datasets, methodologies, and performance evaluation methods used in categorizing dermoscopic lesions. This review focuses on using convolutional neural networks (CNNs) in analyzing PSL. Based on specific inclusion and exclusion criteria, the review included 54 primary studies published on Scopus and PubMed between 2018 and 2022. The results showed that ResNet and self-developed CNN were used in 22% of the studies, followed by Ensemble at 20% and DenseNet at 9%. Public datasets such as ISIC 2019 were predominantly used, and 85% of the classifiers used were softmax. The findings suggest that the input, architecture, and output/feature modifications can enhance the model's performance, although improving sensitivity in multiclass classification remains a challenge. While there is no specific model approach to solve the problem in this area, we recommend simultaneously modifying the three clusters to improve the model's performance.

1. Introduction

Pigmented Skin Lesion (PSL) refers to an area of the skin with abnormal pigmentation compared to the surrounding area [1]. In comparison, melanocytic lesions are commonly observed in pigmented skin lesions. However, pigmentation can also occur in non-melanocytic lesions, such as keratinocyte, vascular, and reactive skin lesions, especially in people with dark skin.

Melanoma is a serious skin cancer that can significantly impact life expectancy. Therefore, detecting melanoma at an early stage may reduce mortality rates [2]–[5]. Melanocytes that undergo cancer cell transformation are the source of melanoma [6]. According to the World Health Organization (WHO), skin cancer has increased over the past few decades; an estimated 2 and 3 million cases are identified globally each year [7]. According to data from 2010 to 2014, Australia has the most significant incidence of melanoma, particularly in Queensland, where 572 cases per 100,000 people are annually [8]. The American Cancer Society has also reported a significant increase in melanoma cases, particularly among older individuals. Between 2011 and 2015, there was an average annual increase of 1.8 % for men and

3.7 % for women in melanoma incidence. [9]. Given the cumulative trend of annual melanoma incidence, researchers are motivated to identify melanoma early to reduce the mortality rates.

In order to treat skin cancer, dermatologists have two main imaging options: dermoscopy or histological imaging obtained through biopsies. Dermoscopy has gained popularity due to its advantages [6], [10]. Histopathological imaging, on the other hand, is an invasive method that requires analysis by the pathology anatomy department, making it costly and time-consuming. Dermatologists frequently use dermoscopic pictures since they are non-invasive, inexpensive, and straightforward. However, dermoscopic images of lesions show a wide range in size, color, and shape. The lack of uniformity in colors, methods, and image capture settings further contributes to the variability of results. Identifying and categorizing dermoscopic images using digital image analysis becomes more challenging due to the presence of various artefacts such as blood vessels, skin hair, dark corners, ink marks, gel bubbles, color charts, and ruler marks [6], [10], [11]. The presence of artefacts and the varying condition and quality of the dermoscopy images can be seen in Fig. 1. These factors, along with the absence of a direct physician investigation, make it difficult for researchers to find a solution to this issue by using various strategies or methodologies that would yield the most significant detection results.

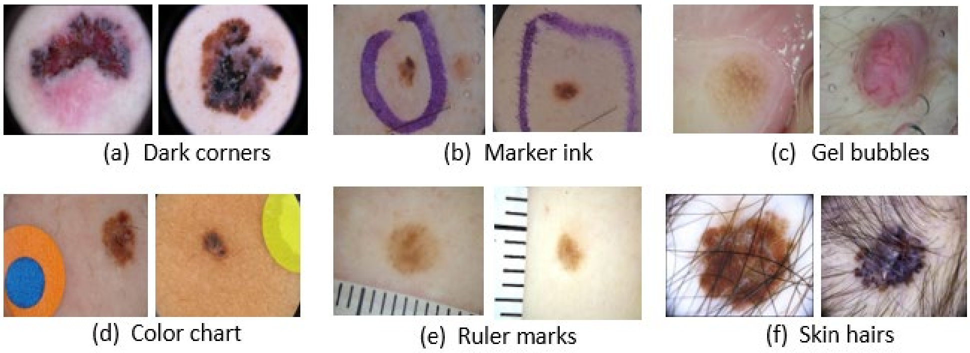

Fig. 1. The condition and quality of the dermoscopy image with the artefact on dataset ISIC 2019.

Research on pigmented skin lesions has rapidly developed in the last five years, mainly using CNN. This condition has motivated mapping existing research to find comprehensive knowledge in the literature. This paper aims to employ using CNN to detect and evaluate the research direction, data sets, and methods applied in pigmented skin lesion research between 2018 and 2022. By conducting a comprehensive literature review, this research aims to facilitate researchers in identifying the latest and unresolved research gaps, which is the main contribution of this paper. This article's structure is described below. Section three presents the research methods used, while section 4 offers responses to research questions, followed by a conclusion.

2. Related Work

The pigmented skin lesion classification study has been relevant in recent decades. It is necessary to consolidate research results to see the direction and trend of research development.

In their research, Pathan *et al.*[12] explored different methodologies and algorithms for the computer-assisted diagnosis of PSL. Similar reviews were conducted by Oliveira *et al.* [13] and Okur and Turkan [6]. This research aims to compile research findings related to the computer image analysis pipeline, including picture pre-processing, segmentation, feature extraction, selection, and classification. Brinker *et al.* [14] did a comprehensive assessment of the categorization of skin cancer with a CNN. However, this study did not specify the review duration, inclusion, and exclusion criteria. Additionally, the systematic review did not employ a validated methodology; therefore, the results may be biased.

The latest review was conducted by Adegun and Viriri [15] and focused on the state-of-the-art assessment of melanoma recognition. This study found that Ensemble deep learning models perform good classification, particularly in pre-processing and segmentation. However, this survey also does not use a proven protocol and only accommodates the outcomes of the ISIC 2018 and ISIC 2019 competitions.

This research aims to conduct a comprehensive review of research related to PSL with CNN in the last five years. This research has advantages over previous similar research on systematic, proven research stages and refers to the systematic literature review published by Kitchenham [16].

3. Method

A systematic examination is a commonly used approach in computer science for reviewing the research literature on pigmented skin lesions. A systematic literature review (SLR) follows a defined research strategy, establishes specific objectives, sets inclusion and exclusion criteria, and creates a qualitative analysis of the articles [17]. In this review, the systematic literature review protocol published by Kitchenhanm [16] was referred to and improved with SLR phases by Jesson *et al.* [17]. Additionally, the approach, technique, and visual representations discussed in this section were incorporated based on the work of Wahono [18].

As indicated in Fig. 2(a), SLR consists of three phases: planning, conducting, and reporting. The introduction highlights that the initial step in the SLR process is to establish the goal of the systematic review. In addition, a review methodology was developed before the review to ensure that any potential biases from the research were. The review approach included research objectives, a search strategy, and an inclusion and exclusion criterion-based selection procedure. The subsequent processes include quality evaluation, data extraction, and synthesis. The review procedure is prepared in sequential order, and during the conducting and reporting phases, an effort will be made to establish, evaluate, and enhance the review methodology as necessary.

3.1. Research Questions

The list of research questions (RQ) consists of inquiries that help narrow the scope of the review process. These questions were formulated based on PICOC criteria (Population, Intervention, Comparison, Outcomes, and Context) [19]. Table 1 contains the PICOC criteria of the research questions. Table 2 provides an overview of the research questions and the underlying motivation that were addressed during the focused process of this review.

Table 1. Synopsis of PICOC

PICOC	Scope
Population	Dermoscopic PSL.
Intervention	The processes of detection, classification, prediction, and analysis are vital components of the study.
Comparison	The method used, the dataset employed, and the comparison of performance are essential aspects considered in the study.
Outcome	The model/method approaches, and the evaluation of their performance are significant factors considered in the study.
Context	Utilizing CNN researches images of PSL.

Table 2. Research questions (RQ) on SLR.

ID	Research Questions	Objectives
RQ1	What kind of CNN is used to analyze PSL data?	Identify CNN's PSL potential and trends.
RQ2	What sets of data are analyzed using PSL?	Find the most popular dataset.
RQ3	What sort of classifiers is the PSL classification using?	Find the dominant classifier.
RQ4	How do the performance results compare to the PSL analysis's methodologies?	Determine the performance outcomes from the PSL analysis methodology.
RQ5	What are performance evaluation matrices employed?	Determine the performance outcomes from the PSL analysis methodology.
RQ6	What unique techniques did the researchers add to CNN?	Identify the uniqueness of the method used by researchers.

The primary studies included in this cover various aspects such as research trends, datasets, techniques, and assessment matrices related to the research questions RQ1 to RQ6. RQ1 focuses on using CNN to analyse pigmented skin lesions, and numerous summaries will be provided. RQ2 examines the commonly used dataset in segmentation, feature extraction, and classification sub-areas. The classifier employed in the classification process will be presented in RQ3. RQ4 provides an overview of the techniques used to examine PSL and their comparative performance. RQ5 highlights the elements considered when assessing the effectiveness of these methods. Finally, RQ6 reveals the distinctiveness of the researchers' approach.

3.2. Search Strategy

The search process involves selecting a digital database, defining the search string, executing and refining the search string, and determining the digital database's main study that matches the search string, as shown in Fig. 2 (b). Before starting the search; it is necessary to determine the commonly used digital databases for relevant research publications. For this SLR, the Scopus and PubMed digital databases were used. A research string was compiled as a search string to facilitate the following steps:

- When selecting search keywords for the PICOC framework, defining the target population and intervention components is essential.
- Formulate search terms that align with the research question.
- Recognize appropriate search terms by examining titles, abstracts, and keywords.
- Recognize alternating spellings and synonyms for the search phrases.
- Construct a search string by combining general search terms using Boolean operators such as “AND” and “OR”.

In the database search, the researchers used "pigmented skin lesion" or "melanoma" in our database search's title, keyword, and abstract. We restricted the search to articles published between the years 2018 and 2022. Additionally, we limited our search to English-language articles published in engineering, computer sciences, mathematics, and medicine. Specifically, the researchers included journal articles and professional conference proceedings

3.3. Study Selection

The selection process starts with defining search strings, executing and refining search strings, retrieving articles from Scopus and PubMed, downloading, and applying to exclude/include criteria shown in Table 3. Initially, the researchers collected 2184 articles based on the search string, but only 1920 articles were accessible. Subsequently, after screening based on the title and abstract, we obtained 409 articles. Among these, 54 were the primary studies for making SLRs based on full text. The researchers consider the quality of the studies as a consequence of research questions, inter-study questions, and similar studies that the authors omitted. The complete list of studies selected is available in Table 5.

Table 3. Inclusion and exclusion criteria.

Inclusion criteria	Inclusion criteria
Scopus and Pubmed include indexes for technical publications.	The writing was not in English.
If duplication is found, we will only include the most recent and comprehensive ones.	Without robust validation and evaluation, it was investigated.
We would include only the journal version rather than proceeding if both existed.	The field of computer image processing does not include studies on dermoscopic pictures.
Just modal dermoscopic pictures.	

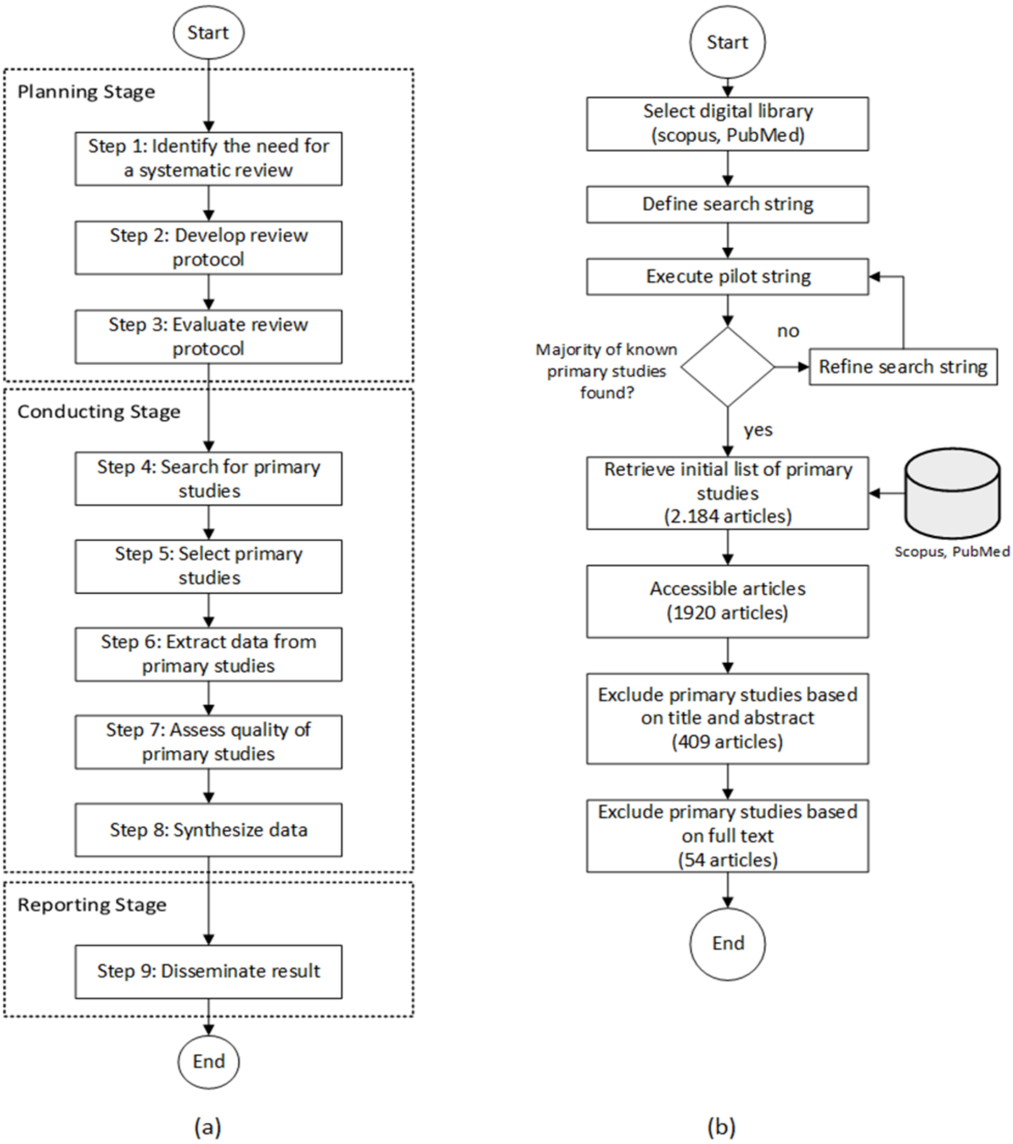

Fig. 2. (a) SLR steps [18], (b) Search and selection process

4. Results and Discussion

The researchers initially collected 2184 articles based on the search string; however, only 1920 articles were accessible. Subsequently, based on the title and abstract, the researchers obtained 409 articles. Of these, 54 were the main study of making SLRs based on full text. The search results in the form of 54 selected papers are then processed to answer the research question (RQ) defined in the PICOC.

4.1. Architecture Model of CNN (RQ1)

Alex-Net is considered a milestone in deep learning architecture [60] as it emerged after winning the 2012 ILSVRC. This triumph has spurred the analysis of pigmented skin lesions using machine learning, particularly deep learning (DL). Over the past decade, researchers have mainly used artificial intelligence (AI), markedly those based on machine learning (ML), to support the analysis of dermoscopic images.

The predominant use of deep learning architecture consists of pre-trained architecture, which usually has proven performance. Pigmented skin lesion analysis using machine learning, especially using CNN, has emerged as the dominant approach. Fig. 3 (a) shows a CNN-based deep learning architecture in this SLR. It was observed that researchers mainly used ResNet and self-develop CNN to analyse PSL, with 22% each, followed by an ensemble with 20%. The utilization of ensemble models that combine several models in a single architecture is currently possible because of the support for fast processing devices.

Vol. 9, No. 3, November 2023, pp. 363-382

However, there are 20% of researchers use their architectural design. The following is a brief explanation of pre-trained models that are widely used by researchers in the classification of PSL:

- The architecture of AlexNet [20] is widely recognized as one of the most influential and prominent CNN architectures in the history of CNN development. It was developed by Krizhevsky, Sutskever, and Geoffrey Hinton in 2012. AlexNet reached significant success by winning the ImageNet LSVRC competition in the same year AlexNet consists of five convolutional layers and three fully connected (FC) layers, with a total parameter count of approximately 60 million. Regarding activation function, ReLU is utilized by AlexNet. The input size for AlexNet is 227x227 pixels.
- The VGG16 [21] architecture is a well-known CNN model recognized for its effectiveness in image classification tasks. It was developed by a team of researchers from the University of Oxford in 2014. VGG16 achieved notable success by winning the ImageNet Large Scale Visual Recognition Challenge competition in the same year. VGG16 consists of 13 convolutional layers and three fully connected layers, with a total parameter count of approximately 138 million. The input size for VGG16 is 224x224 pixels. ReLU serves as the activation function in VGG16.
- ResNet [22] is designed to tackle the degradation problem encountered in deep neural networks. As the depth of the network increases, the accuracy typically reaches a saturation point and rapidly deteriorates. To address this issue, ResNet introduces residual bottleneck blocks. Two types of residual blocks are employed: (1) Identity block and (2) Convolution block. ResNet encompasses several variants, including ResNet50, ResNet50v2, ResNet152, ResNet152v2, and Inception ResNet, each with its architectural specifications and improvements.
- Inception-v3 [23] is an extension of the previous model, achieved by increasing the depth of the network. Several general design principles are used, such as: avoiding representational bottlenecks in early networks, higher dimensional representations, spatial aggregation, and balancing the width and depth of the network. The implementation includes several improvements, such as convolution filter reduction, the use of asymmetric convolutions, improvement of auxiliary classifiers, efficient grid size, and implementation of smoothing regularization.
- DenseNet [24] comprises numerous dense blocks (small convolutional layers, batch normalization, and ReLU Activation). These dense blocks are interconnected through skip connections. Each block is closely interconnected with all the blocks that came before it. A transition layer with batch normalization, 1x1 convolution, and average pooling is introduced between the dense blocks. DenseNet has proven comprehensive exploitation of residual processes while preserving model compactness, leading to competitive precision.
- MobileNet [25] model is based on depth-wise separable convolutions, a factorized convolution technique that divides a standard convolution into a depth-wise convolution and an 11 convolution called a pointwise convolution. Except for the last ultimately linked layer, which is nonlinearity-free and feeds a SoftMax classification layer, all layers are observed by a batch norm and ReLU nonlinearity.
- EfficientNet [26] is a highly advanced architecture renowned for its extensive and robust network design. It achieves an impressive top-1 accuracy of 84.3% on the ImageNet dataset while delivering remarkable efficiency gains. Despite its efficiency, EfficientNet maintains a substantial parameter count of approximately sixty-six million and performs approximately thirty-seven billion floating-point operations per second (FLOPS). EfficientNet is approximately 8.4 times more compact than the best available CNN and provides a 6.1 times faster inference speed.
- The Ensemble is a CNN architecture that usually combines several existing pre-trained CNNs. Of course, this merger has some necessary adjustments according to its purpose. Many researchers have confirmed performance enhancements through the ensemble model. However, it is essential to consider the increased computational requirements associated with this combination.

Pre-trained CNN models must be adjusted to the specific data objects applied as a follow-up. An essential step in this process is adjusting or tuning hyper-parameter values on the chosen model [27]. The success of hyperparameter tuning determines the performance of the model. Poor data adjustment processes and inappropriate hyperparameter tuning may cause bias in the results. The authors recommend using Bayesian tuning, as demonstrated by Swersky *et al.* [28], as an efficient step. The researcher cannot claim the best pre-trained CNN model because the performance results depend on the researcher pre-processing the initial data and tuning the model.

4.2. Use of Dataset (RQ2)

In machine learning research, having access to an appropriate dataset is crucial. Researchers require access to publicly available datasets to replicate earlier studies and conduct their development research, enabling relevant comparisons. The International Skin Imaging Collaboration (ISIC) is a collaborative initiative dedicated to improving melanoma's early detection and accuracy of melanoma to reduce melanoma-related mortality and unnecessary biopsies. The ISIC actively develops recommended dermoscopic standards and provides open access to medical and dermoscopic images of PSL. Fig. 3(b) illustrates the dataset distribution utilized in the study focused on pigmented skin lesions.

Replicating the findings of certain researchers can pose a challenge for other scholars, as approximately 7% of academics employ proprietary datasets. However, a dataset released by ISIC in 2016, extensively utilized in challenges, is widely acknowledged as the standard dataset among scholars. The development of the ISIC/ISBI dataset is presented in Table 4, which highlights the variations in the number of available features, the number of classes, as well as file types and extensions across the datasets.

Researchers should note that the ISIC 2018 and ISIC 2019 subset testing datasets are not published due to their challenges. This implies that when utilizing these datasets, it is necessary to reserve a portion of the training data for validation and testing purposes. When comparing the performance results of different models, researchers must be cautious of result bias that can arise from variations in dataset separation. It is also important to note that the models have the same number of classes, as comparing the performance of models with a different number of classification classes is not valid.

Table 4. ISIC public dataset.

Name	Types and Extension	Distribution			Classification	
		Train	*Val*	*Test*	*Classification*	*Amount of Data*
ISBI/ ISIC 2016	image (jpeg), ground truth (csv)	900	-	379	(1) benign	727
					(2) malignant	173
ISBI/ ISIC 2017	image (jpeg), superpixel mask (png), ground truth (csv)	2,000	150	600	(1) melanoma	374
					(2) seborrheic keratosis	254
					(3) benign nevus	1,372
ISBI/ ISIC 2018 or HAM 10000	image (jpeg)	10,015	193	1,512	(1) melanoma (MEL)	1,113
					(2) melanocytic nevus (NV)	6,705
					(3) basal cell carcinoma (BCC)	514
					(4) actinic keratosis (AKIEC)	327
					(5) benign keratosis (BKL)	1,099
					(6) dermatofibroma (DF)	115
					(7) vascular lesion (VASC)	142
ISBI/ ISIC 2019 or BCN20000	image (jpeg) metadata (age, sex, loc) ground truth (csv)	25,331	-	8,238	(1) melanoma (MEL)	4,522
					(2) melanocytic nevus (NV)	12,875
					(3) basal cell carcinoma (BCC)	3,323
					(4) actinic keratosis (AKIEC)	867
					(5) benign keratosis (BKL)	2,624
					(6) dermatofibroma (DF)	239
					(7) vascular lesion (VASC)	253
					(8) squamous cell carcinoma (SCC)	628

4.3. Use of Classifier (RQ3)

It is widely recognized that pigmented skin lesions have specific requirements for multiple classification tasks. A classifier is a method used to assign a particular class to the supplied data. In order to identify the class or category of input data, the classifier receives input in the form of features obtained from the previous procedure. When applying classifiers, researchers should consider the input data's peculiarities. In this systematic literature review (SLR), Fig. 3(c) visualizes the distribution of the classifier's performance in classifying pigmented skin lesions.

Early studies, which primarily focus on melanoma detection, frequently employ SVM. SVM initially intended to divide data into two classes, necessitating other approaches to handle problems involving many classes. The selection of a suitable classifier determines the success of the model used. Softmax dominates the prevalent choice of classifier usage, particularly for multiclass requirements. Softmax is an activation function commonly used in multiclass classification. Sofmax has advantages, including interpretable probabilities, output normalisation and sensitivity to score differences. However, softmax also has limitations, including dependence on accurate training data, problems with class rewards and sensitivity to training data outliers.

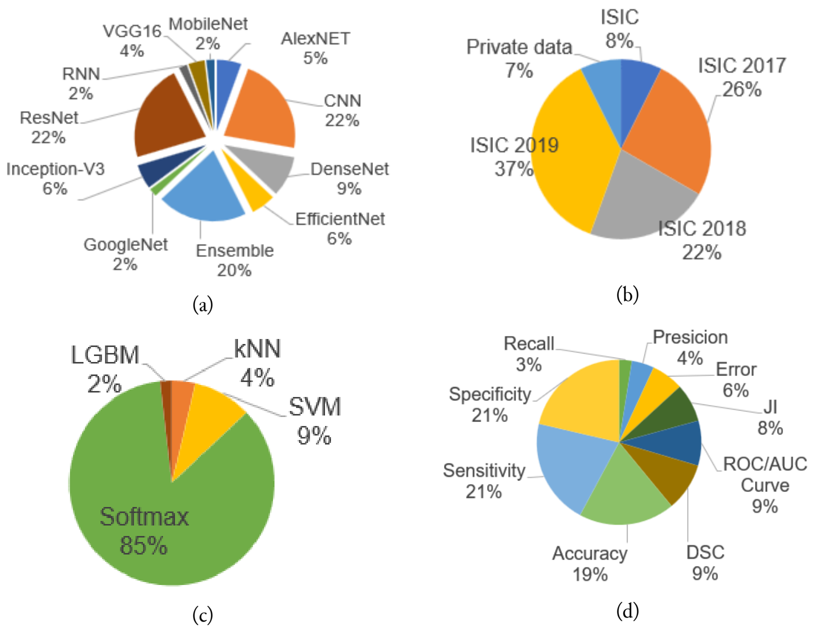

Fig. 3. (a) The use of CNN in the evaluation of PSL; (b) Distribution use of the dataset; (c) Distribution use of classifier; (d) The use of performance evaluation in classifying pigmented skin lesions.

4.4. Performance and Evaluation (RQ4 and RQ5)

Research in pigmented skin lesions, particularly skin cancer diagnosis (melanoma), remains highly motivated by the goal of improving analysis methods. The research focuses on developing increasingly complex deep-learning architectural designs and employing specific methods to enhance performance. Deep learning performance will be improved by creating increasingly complicated deep learning architecture designs with more layers. However, this approach requires significant computational resources, and training takes a while [2]. Research has revealed that deep learning outperforms traditional methods in melanoma detection, mainly when evaluated by skin specialists [83]–[85]. Kassani and Kassani [2] carried out a comparison. With the 2018 ISIC dataset as a comparison, they examine five CNN models for skin pigmentation detection. They found that the accuracy of these models is that of current non-machine learning techniques. Forty-three studies in this SLR were performed, as shown in Table 5. Due to the variations in the dataset, the number of images, and the research aims, evaluating the research performance findings in analyzing pigmented skin lesions presents significant challenges. Obtaining a comparable research performance comparison is particularly difficult, especially for multiclass classifications where the average sensitivity aspect still yields poor results.

Table 5. List of the primary studies based on full-text published paper for SLR

No	Researcher	CNN Model	Augment	Transfer learning	Classifier	Class	Dataset	Unique Method	ACC (%)	SEN (%)	SPE (%)	PRE (%)	AUC (%)
1	Nasiri *et al.* [29]	CNN	×	×	Softmax	2	ISIC	Case-based reasoning	75.0	73.0	78.0	77.0	-
2	Albahar [30]	CNN	×	√	Softmax	2	ISIC	Regularization	97.5	-	-	-	-
3	Abbas and Celebi [31]	RNN	×	×	Softmax	2	ISIC	Features fusion	-	93.0	95.0	-	96.0
4	Hosny, *et al.* [32]	AlexNET	√	√	Softmax	3	ISIC	Transfer learning & augmentation	95.9	88.5	93.0	-	-
5	Hameed, *et al.* [33]	AlexNET	×	×	Softmax	4	Personal data	Multiclass multi-level	96.5	-	-	-	-
6	Haenssle *et al.* [34]	CNN	×	×	Softmax	2	Personal data	-	84.0	95.0	76.7	-	-
7	Kawahara, *et al.* [35]	Inception-V3	×	√	Softmax	5	Personal data	Multimodal Multi-Task Loss Function	-	60.3	78.4	63.3	-
8	Yap *et al.* [36]	ResNet50	×	√	Softmax	5	Personal data	Modality fusion	-	-	-	72.9	86.6
9	Tang *et al.* [37]	CNN	×	×	SVM	3	ISIC2017	Global-part CNN	-	-	-	92.6	-
10	Mahbod, *et al.* [38]	ResNet	×	√	SVM	2	ISIC2017	Multi-scale multi-network	-	-	-	-	87.3
11	Barata and Marques [39]	DenseNet	×	√	Softmax	2	ISIC2017	Hierarchical classification	70.0	-	-	-	-
12	Harangi [40]	Ensemble	√	√	Softmax	3	ISIC2017	Ensemble	-	-	-	-	89.1
13	Yang, *et al.* [41]	ResNet50	×	√	Softmax	3	ISIC2017	Region average pooling	83.0	60.7	88.4	-	84.2
14	Zhang, *et al.* [42]	ResNet	×	√	Softmax	2	ISIC2017	Attention residual learning	85.0	65.8	89.6	-	87.5
15	Yan, *et al.* [43]	VGG16	√	√	Softmax	2	ISIC2017	Visual attention	-	-	-	69.3	85.2
16	Gonzalez-Diaz [44]	ResNet50	√	√	Softmax	3	ISIC2017	Polar average pooling	-	-	-	-	90.8
17	Codella, *et al.* [45]	AlexNet	×	√	kNN	2	ISIC2017	CBIR - hierarchical triplet loss	-	-	-	-	78.6
18	Tschandl, *et al.* [46]	ResNet50	×	×	Softmax	2	ISIC2017	CBIR - cosine similarity	70.8	70.9	74.9	-	81.0
19	Salamaa and Aly [47]	ResNet50	√	√	SVM	2	ISIC2017	-	99.2	98.9	-	98.8	99.3
20	Usmani, *et al.* [48]	CNN	√	√	Softmax	3	ISIC2017	Reinforcement learning	95.3	96.3	98.5	-	-

Table 5. (Continued)

No	Researcher	CNN Model	Augment	Transfer learning	Classifier	Class	Dataset	Unique Method	ACC (%)	SEN (%)	SPE (%)	PRE (%)	AUC (%)
21	Song, *et al.* [49]	Ensemble	√	√	Softmax	3	ISIC2017	Ensemble	90.9	80.8	-	85.9	91.1
22	Wei, *et al.* [50]	CNN	×	√	Softmax	3	ISIC 2017	-	86.3	69.3	90.5	69.4	88
23	Mahbod, *et al.* [51]	EfficientNET	√	√	Softmax	7	ISIC2018	Multi-scale multi-network	88.2	-	-	-	-
24	Almaraz-Damian, *et al.* [52]	CNN	×	√	SVM	2	ISIC2018	Features fusion	-	-	-	-	-
25	Harangi, *et al.* [53]	Inception-V3	×	×	Softmax	7	ISIC2018	Multiclass with binary	67.7	-	-	-	-
26	Barata, *et al.* [54]	Ensemble	×	√	Softmax	2	ISIC2018	Hierarchical attention CNN-LSTM	64.1	63.7	95.6	-	93.9
27	Wu, *et al.* [55]	DenseNet	×	√	Softmax	7	ISIC2018	Multi-input with an attention module	88.4	76.7	96.3	-	97.5
28	Gessert, *et al.*[56]	Ensemble	×	√	Softmax	7	ISIC2018	Patch-based attention	-	67.8	93.1	-	-
29	Maron, *et al.* [57]	ResNet50	√	√	Softmax	5	ISIC2018	-	-	74.4	59.8	-	-
30	Thurnhofer-Hemsi, *et al.* [58]	Ensemble	√	√	Softmax	7	ISIC2018	Regularly spaced shifting	83.2	64.4	95.3	76.1	-
31	Adegun and Viriri [59]	DenseNet	√	√	Softmax	7	ISIC2018	Conditional Random Field (CRF)	98.3	98.5	-	98.0	99.0
32	Pratiwi, *et al.*[60]	Ensemble	√	√	Softmax	7	ISIC2018	Average probability	97.2	90.1	97.7	-	-
33	Rahman and Ami [61]	Ensemble	√	√	Softmax	7	ISIC2018	Ensemble	85.8	-	97.0	81.0	94.0
34	Gutierrez, et.al.[62]	CNN	×	×	Softmax	7	ISIC2018	Statistical fractal signature	88	41	92	46	-
35	Alizadeh and Mahloojifar [63]	CNN	×	√	Softmax	2	ISIC2019	Combining with texture feature	96.7	96.3	97.1	95.1	-
36	Mijwil [64]	Inception-V3	×	√	Softmax	2	ISIC2019	-	86.9	86.1	87.7	87.5	-
37	Noshad, *et al.*[65]	ResNet	×	×	Softmax	2	ISIC2019	The difference in Gaussian feature enhancement	81.0	93.0	-	-	81.5
38	Gessert, *et al.* [66]	Ensemble	√	√	Softmax	8	ISIC2019	Ensemble, multi-input resolution	-	50.9	98.1	-	93.9

Table 5. (Continued)

No	Researcher	CNN Model	Augment	Transfer learning	Classifier	Class	Dataset	Unique Method	ACC (%)	SEN (%)	SPE (%)	PRE (%)	AUC (%)
39	Molina-Molina, *et al.*[67]	DenseNet	×	√	SVM	8	ISIC2019	Features fusion	97.4	66.5	97.9	91.6	-
40	Iqbal, *et al.* [68]	CNN	√	√	Softmax	8	ISIC2019	-	89.6	89.6	97.6	90.6	99.1
41	Putra, *et al.*[69]	EfficientNet	√	√	Softmax	8	ISIC2019	Dynamic training and augment	95.0	65.0	-	-	91.0
42	Kassem, *et al.*[70]	GoogleNet	√	√	Softmax	8	ISIC2019	-	94.9	79.8	97.0	80.4	-
43	Gong, *et al.* [71]	Ensemble	√	√	Softmax	8	ISIC2019	Ensemble	92.4	48.3	97.7	-	91.9
44	Bayram, *et al.*[72]	ResNet	√	√	Softmax	3	ISIC 2019	-	91.2	-	-	-	-
45	Naeem, *et al.*[73]	VGG16	√	√	Softmax	4	ISIC 2019	VGG16 modified	96.9	92.18	-	92.19	-
46	Reis, *et al.*[74]	CNN	×	×	Softmax	2	ISIC 2019	segmentation	91.89	-	-	-	-
47	Nigar, *et al.*[75]	ResNet	×	√	Softmax	8	ISIC 2019	added LIME	94.47	94.01	-	93.57	-
48	Jaisakthi, *et al.*[76]	EfficientNet	√	√	LGBM	2	ISIC 2019	ensemble with metadata	-	-	-	-	96.81
49	Benyahia, *et al.*[77]	DenseNet	√	√	kNN	8	ISIC 2019	Multi-features extraction	92.34	92.75	96.38	85.22	-
50	Filipescu, *et al.*[78]	Resnet	×	√	Softmax	8	ISIC 2019	-	78.1	-	-	-	-
51	Cauvery, *et al.*[79]	Ensemble	×	√	Softmax	8	ISIC 2019	Ensemble	81.2	62	98	73	-
52	Santos, *et al.*[80]	Ensemble	√	√	Softmax	8	ISIC 2019	Ensemble	84.6	-	-	-	-
53	Balaha and Hassan [81]	MobileNet	√	√	Softmax	2	ISIC 2019	sparrow search algorithm	98.27	-	-	-	-
54	Liu, *et al.*[82]	CNN	×	×	Softmax	8	ISIC 2019	Clinical-inspired net	84.9	-	-	54.1	88.2

Most of the journal's categorization performance in this SLR list is evaluated using this section's performance evaluation. The assessment parameters used by the researchers are shown in Fig. 3(b). These factors are commonly used to understand the specifics of classification performance, even though they are not utilized in isolation.

Sensitivity (SEN), also known as recall or true positive rate (TPR), measures the ability of a diagnostic test to identify individuals who have the disease accurately. On the other hand, specificity (SPE), also known as selectivity or true negative rate (TNR), reflects the test's ability to detect entities who do not have the disease accurately. The specificity is determined based on the number of individuals without the condition [63]. The consideration of sensitivity is influenced by the prevalence of the disease rather than the entire population.

Precision refers to the reliability of independent tests conducted under specific conditions. Accuracy (ACC) represents the level of agreement between the obtained quantity or measurement and the true value of the measured quantity. Precision (PREC) quantifies the ratio of true positive predictions to all positive predictions. The following formula can be used to calculate precision:

$$SEN = \frac{TP}{TP+FN} \tag{1}$$

$$SPE = \frac{TN}{TN+FP} \tag{2}$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$PREC = \frac{TP}{TP+FP} \tag{4}$$

“TP” is a symbol for true positives, “TN” is a symbol for true negatives, “FP” is a symbol of false positives, and “FN” is a symbol of false negatives in the confusion matrix.

In cases where the dataset has a significant number of false negatives and a relatively balanced proportion of false positives, selecting a good-accuracy model is advisable. If prioritizing false positives over false negatives is preferred, opting for a sensitive algorithm is recommended. On the other hand, if the objective is to maximize true positives while minimizing false positives, selecting an algorithm with high precision is appropriate. Lastly, choosing an algorithm with high specificity is suitable if the goal is to avoid false positives altogether.

Receiver Operating Characteristics (ROC) is a practical assessment method used to identify the threshold of a model for categorization problems. Evaluating models becomes more convenient with the help of the Area Under the Curve (AUC). The AUC represents the integral or area under the ROC curve. Selecting the model with the highest AUC is essential, as it indicates a greater TP and/or a lower FP at each point. A favourable classification performance is indicated when the resulting curve is closer to the point (0,1). Conversely, if the curve is closer to the baseline line, it indicates a lower classification performance.

Jaccard introduced the Jaccard index (JI) as a numerical measure for comparing or contrasting two sample data objects. The Jaccard index provides a straightforward and intuitive similarity metric between data samples. It is calculated as the ratio of the intersection size to the union size of the two data samples. Similarly, the Dice Similarity Coefficient (DSC) is an index calculated as twice the number of shared elements between two sets divided by the sum of the element counts in each set. JI and DSC are commonly used to assess the effectiveness of segmentation techniques.

Comparing performance outcomes can be challenging when using various datasets, especially during the testing phase. Despite their limitations, many researchers rely on sensitivity and specificity as evaluation measures. Accuracy and precision are less commonly used due to their dependence on threshold values and lack of unique descriptions, as highlighted by Metz [86]. In contrast, the ROC curve provides a comprehensive visualization of all potential classification thresholds, making it a more informative statistic than the error rate classifier typically used by researchers. ROC and AUC are commonly utilized to evaluate classifiers as they offer a more comprehensive assessment. ROC curves remain valuable even when the estimated probabilities are not perfectly calibrated, or the class distribution is imbalanced. Furthermore, ROC curves can be extended to address multiclass scenarios.

4.5. Unique Methods (RQ6)

The researchers employed unique methods to enhance the classification performance. Many researchers have opted for the fully connected network (FCN) classifier with the softmax activation function. Additionally, most researchers apply transfer learning by utilizing pre-trained weights to expedite the training process. Some researchers also apply augmentation to address the class imbalance. Researchers widely utilize the ISIC dataset due to its comprehensive features provided to support research. Standard classification performance parameters include accuracy, sensitivity, specificity, precision, and AUC.

Several researchers employ novel approaches to enhance the performance of their models by making fundamental design modifications. Incorporating features obtained by handcrafted and CNN output in features fusion improves classification performance [31], [51], [52], [67]. Furthermore, the combination of inputs in multi-modalities and images with different resolutions to obtain global and local features is also widely carried out, such as research [33], [35], [36], [51], [55], [56], [87]. This combination of inputs also prompts researchers to use a variety of models (Ensemble).

Several researchers, such as Barata *et al.* [39] and Codella *et al.* [45], have researched the interpretability of models, aiming to emulate the process of medical examinations performed by dermatologists by applying a hierarchy to their classification. However, this hierarchical concept has weaknesses, including lengthy processing time and resource consumption, while not aiding in interpreting classification results since the process remains hidden from the user. In contrast, Zhang *et al.* [42] and Yan *et al.* [43] employ an attention map to interpret the classification results by highlighting the image part that contributes the most to the decision results. Although an attention map is claimed to increase the interpretability of the model, these two studies were only carried out in binary classification. Tschandl *et al.* [46] and Codella *et al.* [45] used content-based image retrieval (CBIR) to develop interpretable machine learning (IML) in the classification of PSL. The unique methods introduced by these researchers can be categorized into three groups: input/modalities, output/features, and architectures, as shown in Table 6.

Table 6. A unique method was added.

Area	Unique Method
Input/modalities	Augmentation [32], [69]
	Segmentation [74]
	Multiclass multi-level [33]
	Multimodal multi-task [35]
	Modality fusion [36]
	Multiscale multi-network [38], [51]
	Multi-input [41], [55]
Architectures	Case-based reasoning (CBR) [29]
	Regularization [30]
	Hierarchical [38], [39], [43], [54]
	Ensemble [40], [49], [54], [56], [58], [60], [61], [66], [71], [76], [79], [80]
	Pooling [28], [41], [35]
	Attention [42], [43], [54]–[56]
	Triplet loss [45]
Output/Features	Features fusion [22], [43], [54], [56], [58], [77]
	Global-part CNN [34]
	Multi-output [66]

4.6. Implications for Clinical Practice and Future Direction

The application of CNN in PSL classification is a computer-aided diagnostic tool designed to help doctors, primarily dermatologists [10]. Based on Table 5, the classification performance shows consistent improvement as the number of disease classes that expansion with the availability of the dataset. In some cases, the model performance results have surpassed the relevant doctors for some cases [85], but it still leaves challenges related to model interpretability. The medical field places a strong emphasis on interpretability for diagnostic aids. At the clinical implementation stage, it needs to be down streamed by involving relevant medical personnel to suit their needs. Based on this literature review, future research aims to further enhance the model performance, especially for multi-class classification and the interpretability challenges associated with their application in the medical field.

5. Conclusion

This systematic literature review (SLR) analyses the trends, datasets, methodologies, and evaluation matrices employed in studying pigmented skin lesions from 2018 to 2022, employing convolutional neural networks (CNN). Fifty-four papers that met the inclusion and exclusion criteria were examined. Personal datasets were utilized in 7% of the research, but the ISIC dataset emerged as the dominant choice due to its extensive data and continuous growth. Pre-trained models and transfer learning were commonly adopted for their performance benefits. Classifiers such as softmax-based and support vector machines (SVM) were widely utilized. Researchers frequently made modifications within three main clusters to enhance model performance. Although no specific model approach universally solves the problem in this field, it is recommended to simultaneously modify the three clusters to improve the overall model performance. Model interpretability has emerged as a research direction with potential real-world applications.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. This research is funded by LPDP (Indonesia Endowment Fund for Education).
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] A. Hon and A. Oakley, "Pigmented skin lesions," *DermNet NZ*, 2015. Available at: https://dermnetnz.org/topics/pigmented-skin-lesions

[2] S. H. Kassani and P. H. Kassani, "A comparative study of deep learning architectures on melanoma detection," *Tissue Cell*, vol. 58, no. April, pp. 76–83, 2019, doi: 10.1016/j.tice.2019.04.009.

[3] A. M. Alqudah, H. Alquran, and I. A. Qasmieh, "Segmented and non-segmented skin lesions classification using transfer learning and adaptive moment learning rate technique using pretrained convolutional neural network," *J. Biomimetics, Biomater. Biomed. Eng.*, vol. 42, pp. 67–78, 2019, doi: 10.4028/www.scientific.net/JBBBE.42.67.

[4] M. Cullell-Dalmau, M. Otero-Viñas, and C. Manzo, "Research Techniques Made Simple: Deep Learning for the Classification of Dermatological Images," *J. Invest. Dermatol.*, vol. 140, no. 3, pp. 507–514, Mar. 2020, doi: 10.1016/j.jid.2019.12.029.

[5] A. Jibhakate, P. Parnerkar, S. Mondal, V. Bharambe, and S. Mantri, "Skin Lesion Classification using Deep Learning and Image Processing," in *3rd International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, Dec. 2020, pp. 333–340. doi: 10.1109/ICISS49785.2020.9316092.

[6] E. Okur and M. Turkan, "A survey on automated melanoma detection," *Eng. Appl. Artif. Intell.*,

vol. 73, no. March, pp. 50–67, Aug. 2018, doi: 10.1016/j.engappai.2018.04.028.

[7] WHO, "Skin Cancers," 2019. Available at: https://www.iarc.who.int/cancer-type/skin-cancer/.

[8] S. A. Margolis, "Skin cancer medicine integral to Australian general practice," *Aust. J. Gen. Pract.*, vol. 48, no. 6, pp. 343–343, Jun. 2019, doi: 10.31128/AJGP-06-19-1234e.

[9] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA. Cancer J. Clin.*, vol. 69, no. 1, pp. 7–34, Jan. 2019, doi: 10.3322/caac.21551.

[10] M. E. Celebi, N. Codella, and A. Halpern, "Dermoscopy Image Analysis: Overview and Future Directions," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 2, pp. 474–478, 2019, doi: 10.1109/JBHI.2019.2895803.

[11] M. H. Jafari, N. Karimi, E. Nasr-Esfahani, S. Samavi, and S. M. R. Soroushmehr, "Skin lesion segmentation in clinical images using deep learning," in *23rd International Conference on Pattern Recognition (ICPR)*, IEEE, Dec. 2016, pp. 337–342. doi: 10.1109/ICPR.2016.7899656.

[12] S. Pathan, K. G. G. Prabhu, and P. C. C. Siddalingaswamy, "Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—A review," *Biomed. Signal Process. Control*, vol. 39, pp. 237–262, 2018, doi: 10.1016/j.bspc.2017.07.010.

[13] R. B. Oliveira, J. P. Papa, A. S. Pereira, and J. M. R. S. Tavares, "Computational methods for pigmented skin lesion classification in images: review and future trends," *Neural Comput. Appl.*, vol. 29, no. 3, pp. 613–636, 2018, doi: 10.1007/s00521-016-2482-6.

[14] T. J. Brinker, A. Hekler, J. S. Utikal, N. Grabe, D. Schadendorf, and J. Klode, "Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review," *J. Med. Internet Res.*, vol. 20, no. 10, pp. 1–8, Oct. 2018, doi: 10.2196/11936.

[15] A. Adegun and S. Viriri, "Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art," *Artif. Intell. Rev.*, vol. 54, no. 2, pp. 811–841, Feb. 2021, doi: 10.1007/s10462-020-09865-y.

[16] B. Kitchenham, *Procedures for Performing Systematic Reviews, Version 1.1*. Keele University Technical Report TR/SE-0401, 2005. Available at: https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf

[17] J. Jesson, L. Matheson, and F. M. Lacey, *Doing your systematic review - Traditional and systematic techniques*. SAGE Publications Ltd, 2011. Available at: https://www.worldcat.org/title/706789667

[18] R. S. Wahono, "A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks," *J. Softw. Eng.*, vol. 1, no. 1, pp. 1–12, Jan. 2015. Available at: https://media.neliti.com/media/publications/90270-EN-a-systematic-literature-review-of-softwa

[19] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," EBSE Technical Report Version 2.3 - University of Durham, 2007. Available at: https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *in a conference paper at ICLR 2015,* pp. 1–14, 2014. doi: 10.48550/arXiv.1409.1556.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 2818–2826, Dec. 2016, doi: 10.1109/CVPR.2016.308.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, Nov. 2017, doi: 10.1109/CVPR.2017.243.

[25] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," pp. 1-9, doi: 10.48550/arXiv.1704.04861.

[26] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *36th Int. Conf. Mach. Learn. ICML 2019*, pp. 10691–10700, 2019. doi: 10.48550/arXiv.1905.11946.

[27] H. Cho, Y. Kim, E. Lee, D. Choi, Y. Lee, and W. Rhee, "Basic Enhancement Strategies When Using Bayesian Optimization for Hyperparameter Tuning of Deep Neural Networks," *IEEE Access*, vol. 8, pp. 52588–52608, 2020, doi: 10.1109/ACCESS.2020.2981072.

[28] K. Swersky, J. Snoek, and R. P. Adams, "Multi-task Bayesian optimization," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2013. Available at: http://papers.neurips.cc/paper/5086-multi-task-bayesian-optimization.pdf.

[29] S. Nasiri, J. Helsper, M. Jung, and M. Fathi, "DePicT Melanoma Deep-CLASS: A deep convolutional neural networks approach to classify skin lesion images," *BMC Bioinformatics*, vol. 21, pp. 1–13, 2020, doi: 10.1186/s12859-020-3351-y.

[30] M. A. Albahar, "Skin Lesion Classification Using Convolutional Neural Network With Novel Regularizer," *IEEE Access*, vol. 7, pp. 38306–38313, 2019, doi: 10.1109/ACCESS.2019.2906241.

[31] Q. Abbas and M. E. Celebi, "DermoDeep-A classification of melanoma-nevus skin lesions using multi-feature fusion of visual features and deep neural network," *Multimed. Tools Appl.*, vol. 78, no. 16, pp. 23559–23580, 2019, doi: 10.1007/s11042-019-7652-y.

[32] K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Classification of skin lesions using transfer learning and augmentation with Alex-net," *PLoS One*, vol. 14, no. 5, pp. 1-17 May 2019, doi: 10.1371/journal.pone.0217293.

[33] N. Hameed, A. M. Shabut, M. K. Ghosh, and M. A. Hossain, "Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques," *Expert Syst. Appl.*, vol. 141, pp. 1–18, Mar. 2020, doi: 10.1016/j.eswa.2019.112961.

[34] H. A. Haenssle *et al.*, "Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions," *Ann. Oncol.*, vol. 31, no. 1, pp. 137–143, Jan. 2020, doi: 10.1016/j.annonc.2019.10.013.

[35] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 2, pp. 538–546, Mar. 2019, doi: 10.1109/JBHI.2018.2824327.

[36] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Exp. Dermatol.*, vol. 27, no. 11, pp. 1261–1267, Nov. 2018, doi: 10.1111/exd.13777.

[37] P. Tang, Q. Liang, X. Yan, S. Xiang, and D. Zhang, "GP-CNN-DTEL: Global-Part CNN Model With Data-Transformed Ensemble Learning for Skin Lesion Classification," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 10, pp. 2870–2882, Oct. 2020, doi: 10.1109/JBHI.2020.2977013.

[38] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang, "Fusing fine-tuned deep features for skin lesion classification," *Comput. Med. Imaging Graph.*, vol. 71, pp. 19–29, Jan. 2019, doi: 10.1016/j.compmedimag.2018.10.007.

[39] C. Barata and J. S. Marques, "Deep Learning For Skin Cancer Diagnosis With Hierarchical Architectures," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, Apr. 2019, pp. 841–845. doi: 10.1109/ISBI.2019.8759561.

[40] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," *J. Biomed. Inform.*, vol. 86, no. August, pp. 25–32, Oct. 2018, doi: 10.1016/j.jbi.2018.08.006.

[41] J. Yang, F. Xie, H. Fan, Z. Jiang, and J. Liu, "Classification for Dermoscopy Images Using Convolutional Neural Networks Based on Region Average Pooling," *IEEE Access*, vol. 6, pp. 65130–65138, 2018, doi: 10.1109/ACCESS.2018.2877587.

[42] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention Residual Learning for Skin Lesion Classification," *IEEE Trans. Med. Imaging*, vol. 38, no. 9, pp. 2092–2103, Sep. 2019, doi: 10.1109/TMI.2019.2893944.

[43] Y. Yan, J. Kawahara, and G. Hamarneh, "Melanoma Recognition via Visual Attention," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer International Publishing, 2019, pp. 793–804. doi: 10.1007/978-3-030-20351-1_62.

[44] I. González-Díaz, "DermaKNet: Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for Skin Lesion Diagnosis," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 2, pp. 547–559, 2019, doi: 10.1109/JBHI.2018.2806962.

[45] N. C. F. Codella, C.-C. Lin, A. Halpern, M. Hind, R. Feris, and J. R. Smith, "Collaborative Human-AI (CHAI): Evidence-Based Interpretable Melanoma Classification in Dermoscopic Images," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, , vol. 11038 LNCS, pp. 97–105, Springer, 2018. doi: 10.1007/978-3-030-02628-8_11.

[46] P. Tschandl, G. Argenziano, M. Razmara, and J. Yap, "Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features," *Br. J. Dermatol.*, vol. 181, no. 1, pp. 155–165, 2019, doi: 10.1111/bjd.17189.

[47] W. M. Salamaa and M. H. Aly, "Deep learning design for benign and malignant classification of skin lesions: a new approach," *Multimed. Tools Appl.*, vol. 80, no. 17, pp. 26795–26811, Jul. 2021, doi: 10.1007/s11042-021-11000-0.

[48] U. A. Usmani, J. Watada, J. Jaafar, I. A. Aziz, and A. Roy, "A reinforcement learning algorithm for automated detection of skin lesions," *Appl. Sci.*, vol. 11, no. 20, pp.9367, 2021, doi: 10.3390/app11209367.

[49] J. Song, J. Li, S. Ma, J. Tang, and F. Guo, "Melanoma Classification in Dermoscopy Images via Ensemble Learning on Deep Neural Network," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 751–756. doi: 10.1109/BIBM49941.2020.9313451.

[50] Z. Wei, F. Shi, L. Chen, Q. Li, and H. Song, "Multi-level contexts aggregation for melanoma recognition under feature confusion regularization," *Signal, Image Video Process.*, vol. 16, no. 2, pp. 411–418, Mar. 2022, doi: 10.1007/s11760-021-01949-8.

[51] 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," *Comput. Methods Programs Biomed.*, vol. 193, pp. 1–9, 2020, doi: 10.1016/j.cmpb.2020.105475.

[52] J.-A. Almaraz-Damian, V. Ponomaryov, S. Sadovnychiy, and H. Castillejos-Fernandez,

"Melanoma and Nevus Skin Lesion Classification Using Handcraft and Deep Learning Feature Fusion via Mutual Information Measures," *Entropy*, vol. 22, no. 4, p. 484, Apr. 2020, doi: 10.3390/e22040484.

[53] B. Harangi, A. Baran, and A. Hajdu, "Assisted deep learning framework for multi-class skin lesion classification considering a binary classification support," *Biomed. Signal Process. Control*, vol. 62, p. 102041, 2020, doi: 10.1016/j.bspc.2020.102041.

[54] C. Barata, J. S. Marques, and M. E. Celebi, "Deep Attention Model for the Hierarchical Diagnosis of Skin Lesions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2757-2765, 2019. doi: 10.1109/CVPRW.2019.00334.

[55] J. Wu, W. Hu, Y. Wang, and Y. Wen, "A Multi-Input CNNs with Attention for Skin Lesion Classification," in *2020 IEEE International Conference on Smart Cloud (SmartCloud)*, IEEE, Nov. 2020, pp. 78–83. doi: 10.1109/SmartCloud49737.2020.00023.

[56] N. Gessert *et al.*, "Skin Lesion Classification Using CNNs with Patch-Based Attention and Diagnosis-Guided Loss Weighting," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 495–503, 2020, doi: 10.1109/TBME.2019.2915839.

[57] R. C. Maron *et al.*, "Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks," *Eur. J. Cancer*, vol. 119, pp. 57–65, Sep. 2019, doi: 10.1016/j.ejca.2019.06.013.

[58] K. Thurnhofer-Hemsi, E. Lopez-Rubio, E. Dominguez, and D. A. Elizondo, "Skin lesion classification by ensembles of deep convolutional networks and regularly spaced shifting," *IEEE Access*, vol. 9, pp. 112193–112205, 2021, doi: 10.1109/ACCESS.2021.3103410.

[59] A. A. Adegun and S. Viriri, "FCN-Based DenseNet Framework for Automated Detection and Classification of Skin Lesions in Dermoscopy Images," *IEEE Access*, vol. 8, pp. 150377–150396, 2020, doi: 10.1109/ACCESS.2020.3016651.

[60] R. A. Pratiwi, S. Nurmaini, D. P. Rini, M. N. Rachmatullah, and A. Darmawahyuni, "Deep ensemble learning for skin lesions classification with convolutional neural network," *IAES Int. J. Artif. Intell.*, vol. 10, no. 3, pp. 563–570, 2021, doi: 10.11591/ijai.v10.i3.pp563-570.

[61] Z. Rahman and A. M. Ami, "A transfer learning based approach for skin lesion classification from imbalanced data," *Proc. 2020 11th Int. Conf. Electr. Comput. Eng. ICECE 2020*, pp. 65–68, 2020, doi: 10.1109/ICECE51571.2020.9393155.

[62] J. A. Camacho-Gutiérrez, S. Solorza-Calderón, and J. Álvarez-Borrego, "Multi-class skin lesion classification using prism- and segmentation-based fractal signatures," *Expert Syst. Appl.*, vol. 197, no. November 2021, p. 116671, Jul. 2022, doi: 10.1016/j.eswa.2022.116671.

[63] S. M. Alizadeh and A. Mahloojifar, "Automatic skin cancer detection in dermoscopy images by combining convolutional neural networks and texture features," *Int. J. Imaging Syst. Technol.*, vol. 31, no. 2, pp. 695–707, Jun. 2021, doi: 10.1002/ima.22490.

[64] M. M. Mijwil, "Skin cancer disease images classification using deep learning solutions," *Multimed. Tools Appl.*, vol. 80, no. 17, pp. 26255–26271, 2021, doi: 10.1007/s11042-021-10952-7.

[65] A. Noshad, A. Khonaksar, and M. Mohebbi, "SkinXNet: A DoG-based Model for Automatic Detection of Skin Lesion using Deep Learning," in *5th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, IEEE, Apr. 2021, pp. 1–6. doi: 10.1109/IPRIA53572.2021.9483522.

[66] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using

ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, p. 100864, 2020, doi: 10.1016/j.mex.2020.100864.

[67] E. O. Molina-Molina, S. Solorza-Calderón, and J. Álvarez-Borrego, "Classification of Dermoscopy Skin Lesion Color-Images Using Fractal-Deep Learning Features," *Appl. Sci.*, vol. 10, no. 17, p. 5954, Aug. 2020, doi: 10.3390/app10175954.

[68] I. Iqbal, M. Younus, K. Walayat, M. U. Kakar, and J. Ma, "Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images," *Comput. Med. Imaging Graph.*, vol. 88, no. November 2020, p. 101843, Mar. 2021, doi: 10.1016/j.compmedimag.2020.101843.

[69] T. A. Putra, S. I. Rufaida, and J. S. Leu, "Enhanced Skin Condition Prediction through Machine Learning Using Dynamic Training and Testing Augmentation," *IEEE Access*, vol. 8, pp. 40536–40546, 2020, doi: 10.1109/ACCESS.2020.2976045.

[70] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin Lesions Classification Into Eight Classes for ISIC 2019 Using Deep Convolutional Neural Network and Transfer Learning," *IEEE Access*, vol. 8, pp. 114822–114832, 2020, doi: 10.1109/ACCESS.2020.3003890.

[71] A. Gong, X. Yao, and W. Lin, "Classification for Dermoscopy Images Using Convolutional Neural Networks Based on the Ensemble of Individual Advantage and Group Decision," *IEEE Access*, vol. 8, pp. 155337–155351, 2020, doi: 10.1109/ACCESS.2020.3019210.

[72] B. Bayram *et al.*, "Classification of Skin Lesion Images with Deep Learning Approaches," *Balt. J. Mod. Comput.*, vol. 10, no. 2, pp. 241–250, 2022, doi: 10.22364/bjmc.2022.10.2.10.

[73] A. Naeem, M. S. Farooq, A. Khelifi, and A. Abid, "Malignant Melanoma Classification Using Deep Learning: Datasets, Performance Measurements, Challenges and Opportunities," *IEEE Access*, vol. 8, pp. 110575–110597, 2020, doi: 10.1109/ACCESS.2020.3001507.

[74] H. C. Reis, V. Turk, K. Khoshelham, and S. Kaya, "InSiNet: a deep convolutional approach to skin cancer detection and segmentation," *Med. Biol. Eng. Comput.*, vol. 60, no. 3, pp. 643–662, Mar. 2022, doi: 10.1007/s11517-021-02473-0.

[75] N. Nigar, M. Umar, M. K. Shahzad, S. Islam, and D. Abalo, "A Deep Learning Approach Based on Explainable Artificial Intelligence for Skin Lesion Classification," *IEEE Access*, vol. 10, no. November, pp. 113715–113725, 2022, doi: 10.1109/ACCESS.2022.3217217.

[76] J. S M, M. P, C. Aravindan, and R. Appavu, "Classification of skin cancer from dermoscopic images using deep neural network architectures," *Multimed. Tools Appl.*, vol. 82, no. 10, pp. 15763–15778. 2022, doi: 10.1007/s11042-022-13847-3.

[77] S. Benyahia, B. Meftah, and O. Lézoray, "Multi-features extraction based on deep learning for skin lesion classification," *Tissue Cell*, vol. 74, no. November 2021, p. 101701, Feb. 2022, doi: 10.1016/j.tice.2021.101701.

[78] S. Filipescu, A. Butacu, G. Tiplica, and D. Nastac, "Deep-learning approach in the study of skin lesions," *Ski. Res. Technol.*, vol. 27, no. 5, pp. 931–939, Sep. 2021, doi: 10.1111/srt.13045.

[79] Cauvery, Siddalingaswamy, S. Pathan, and N. D'souza, "A Multiclass Skin Lesion classification approach using Transfer learning based convolutional Neural Network," in *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)*, IEEE, Mar. 2021, pp. 1–6. doi: 10.1109/ICBSII51839.2021.9445175.

[80] F. Santos, F. Silva, and P. Georgieva, "Transfer Learning for Skin Lesion Classification using Convolutional Neural Networks," in *International Conference on Innovations in Intelligent SysTems and Applications (INISTA)*, IEEE, Aug. 2021, pp. 1–6. doi: 10.1109/INISTA52262.2021.9548455.

[81] H. M. Balaha and A. E.-S. Hassan, "Skin cancer diagnosis based on deep transfer learning and sparrow search algorithm," *Neural Comput. Appl.*, vol. 9, pp. 815-853, Sep. 2022, doi: 10.1007/s00521-022-07762-9.

[82] Z. Liu, R. Xiong, and T. Jiang, "CI-Net: Clinical-Inspired Network for Automated Skin Lesion Recognition," *IEEE Trans. Med. Imaging*, vol. 42, no. 3, pp. 619–632, Mar. 2023, doi: 10.1109/TMI.2022.3215547.

[83] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, "Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm," *J. Invest. Dermatol.*, vol. 138, no. 7, pp. 1529–1538, Jul. 2018, doi: 10.1016/j.jid.2018.01.028.

[84] Y. Fujisawa *et al.*, "Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis," *Br. J. Dermatol.*, vol. 180, no. 2, pp. 373–381, 2019, doi: 10.1111/bjd.16924.

[85] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[86] C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.*, vol. 8, no. 4, pp. 283–298, Oct. 1978, doi: 10.1016/S0001-2998(78)80014-2.

[87] P. Tang *et al.*, "Efficient skin lesion segmentation using separable-Unet with stochastic weight averaging," *Comput. Methods Programs Biomed.*, vol. 178, pp. 289–301, 2019, doi: 10.1016/j.cmpb.2019.07.005.

[88] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks," *IEEE Trans. Med. Imaging*, vol. 36, no. 4, pp. 994–1004, Apr. 2017, doi: 10.1109/TMI.2016.2642839.