Distributed Rough Set Based Feature Selection Approach to Analyse Deep and Hand-crafted Feature for mammography Mass Classification

Hamidinekoo, Azam; Chelly Dagdia, Zaineb; Suhail, Zobia; Zwiggelaar, Reyer

Published in:
2018 IEEE International Conference on BIG DATA

Publication date:
2018

Citation for published version (APA):
Hamidinekoo, A., Chelly Dagdia, Z., Suhail, Z., & Zwiggelaar, R. (2018). Distributed Rough Set Based Feature Selection Approach to Analyse Deep and Hand-crafted Feature for mammography Mass Classification. In *2018 IEEE International Conference on BIG DATA* IEEE Press.

Distributed Rough Set Based Feature Selection Approach to Analyse Deep and Hand-crafted Features for Mammography Mass Classification

Azam Hamidinekoo*, Zaineb Chelly Dagdia*†, Zobia Suhail*‡, and Reyer Zwiggelaar*
*Department of Computer Science, Aberystwyth University, Aberystwyth, United Kingdom
{azh2, zaineb.chelly, zoa1, rrz}@aber.ac.uk
†LARODEC, Institut Supérieur de Gestion de Tunis, Tunis, Tunisia
chelly.zaineb@gmail.com
‡Punjab University College of Information Technology
zobia.souhail@pucit.edu.pk

***Abstract*—Breast cancer has a high incidence among women worldwide. This, together with the recent developments in deep learning based convolutional networks, have motivated research towards the enhancement of Computer Aided Diagnosis (CAD) systems. In this paper, the performance of a densely connected convolutional network (DenseNet) for breast cancer was investigated for the malignant/benign classification of mammographic masses. Different mammography data sets were collected to investigate the capacity of this network for learning a combination of these databases. To achieve this, internal low-level, mid-level and high-level features/abstracts were extracted from the model together with hand-crafted features, generating a vast amount of data. Using the distributed rough set based feature selection approach (Sp-RST), significant features were selected from both deep learning based features and hand-crafted ones, and fed into a learning model with separate and combined data approaches for the classification of mammographic masses. Results show that by using Sp-RST as a powerful technique capable of performing big data preprocessing, DenseNet had the representational capacity to learn mammographic abnormalities.**

***Index Terms*—Breast Cancer; Feature Selection; DenseNet; Big Data; Mass Classification**

I. Introduction

Breast cancer is the most regularly diagnosed cancer and recent estimates highlight that more than 1 050 000 cancerous cases occur every year in the world, with approximately 580 000 case studies in developed countries and nearly 470 000 in developing countries [1]. Currently, mammography is considered as a common modality used for primary lesion visualisation and detecting early changes in breast tissue. Breast cancer can be visualised in mammograms as masses, architectural distortion, and microcalcifications. Clinicians use several factors to diagnose their potentially cancerous nature. Among these factors are size, morphology, distribution, form, shape, intensity and density of the abnormalities. Computer Aided Diagnosis (CAD) systems have been developed as an alternative to assist radiologists' interpretation and to improve their diagnostic accuracy for the patients' outcomes. These systems aim to improve the identification of subtle suspicious abnormalities in mammograms [2], [3].

Delineation of breast tissue characteristics has been done using traditional machine learning methods [4]. Several hand-crafted (manual) features have been explored for classifying mass abnormalities. The common features include statistical, textural, morphological and intensity based characteristics. Some histogram-based intensity features and Gray Level Co-occurrence Matrix (GLCM) based texture features (contrast, correlation, energy and homogeneity) [2] have also been explored for discriminating cancerous and non-cancerous mammographic masses. Several studies have evaluated the performance of these manual feature sets using several learning methods such as Random Forest (RF), Support Vector Machine (SVM), Decision Tree and Neural Networks [2], [4].

Recently, deep learning models including deep convolutional neural networks (CNNs) [5], inspired by information processing in animal visual cortex, have shown remarkable results in various image processing tasks. The intuition behind deep convolutional models is to learn a hierarchical representation of input data, without relying on hand-crafted features, via a cascade of multiple layers of nonlinear processing units [6]. Accordingly, an automatic feature extraction and transformation is performed from one layer to another in an hierarchic concept. In such a way, the output feature maps from the previous layer are fed as inputs to the successive layer, leading to multiple levels of representations corresponding to different levels of abstraction in the hierarchy of concepts.

Incorporating deep learning concept and methods into the wide range of mammographic applications have expanded ideas to modify CAD systems. Among these, Dhunge *et al.* [7] combined a mass candidate generator and a CNN to define texture and morphology related features for a linear SVM classifier. Carneiro *et al.* [8] did transfer learning with a previously trained CNN and fine-tuned it using unregistered mammograms for the task of mass segmentation. Huynh *et al.* [9] also used transfer learning to extract tumour related information to distinguish between benign/malignant breast lesions. Kooi *et al.* [10] investigated a CAD system relying on manual and CNN designed features. Data augmentation and context effects for classifying pre-segmented masses were studied in [11]. Several CNNs with various depths were evaluated by Arevalo *et al.* [12], comparing the best obtained results from Histogram of Oriented Gradients, Histogram of Gradient Divergence and hand-crafted features. However, none of these studies have investigated the behaviour of the deep learning based methods for the learning process, to express why and how the deep learning based networks perform so well. Therefore, in this paper, we aim to investigate the performance of a densely connected neural network (DenseNet) [13] for breast cancer by analysing the network's intermediate feature maps extracted from different levels of the model's architecture. Accordingly, many features were extracted leading to a vast amount of data that could be referred to as big data. This data was computationally expensive to analyse with standard techniques. To investigate and analyse the DenseNet internal information via the generated feature maps, the use of an appropriate feature selection approach that could preserve the semantics of the features, analyse the facts hidden in data and find a minimal knowledge representation without sacrificing performance of the learning model is essential.

With regards to the feature reduction techniques, and in the context of big data, various distributed approaches have been proposed in literature and these can be grouped into two categories namely methods that perform a transformation on the original meaning of the attributes, named the *transformation based approaches* (also called *feature extraction approaches*), and techniques which preserve the semantic of the features called the *selection based approaches* [14]. Feature extraction generally refers to approaches that build combinations of variables to represent the initial set of attributes. This is achieved via the new set of constructed variables while still representing the data with satisfactory accuracy. Feature extraction approaches are usually employed in cases where the semantics of the initial data set (initial features) will not be required to perform any future actions. On the other side, the selection based approaches aim to retain the semantic (meaning) of the initial attribute set. The major aim of these approaches is to find a minimal sub-set of features from a given problem domain, while retaining a sufficient accuracy in describing the initial features [14]. Feature selection techniques can be further partitioned into *filter approaches* and *wrapper approaches*. The main difference between the two categories is that wrapper approaches include a learning algorithm in the feature sub-set evaluation, and hence they are tied to a particular induction algorithm. In this work, we focus on the application of a feature selection approach, specifically a filter technique. This was important to preserve the meaning/semantics of the DenseNet generated feature maps and to have a better understanding of the model behaviour. However, most of these distributed feature selection approaches require the user to deal with the algorithms' parameterisation, noise levels specification or to give a threshold that decides when the algorithm should end; which are all counted as significant drawbacks. All of these require users to make a decision based on their own (possibly subjective) perception. To overcome these limitations, we used the distributed rough set based feature selection approach (Sp-RST) [15]. Sp-RST, dedicated to big data feature selection, is a distributed implementation design of the standard Rough Set Theory (RST) [16], which is a powerful feature selection technique that has made many achievements in many applications such as in environment, epidemiology, medicine and many others [17], [18].

As mentioned earlier, DenseNet is used to classify mammographic mass abnormalities into benign and malignant classes using various mammographic data sets that were acquired from different laboratories, with various scanners and approaches. Within this application, the main motivations of our paper are to (1) empirically demonstrate the effectiveness of DenseNet for binary classification of mass abnormalities on mammograms, and (2) investigate the behaviour of a deep and dense convolutional model for the learning process through the generated deep features/abstracts. To achieve this, after creating a pool of deep feature maps and in order to evaluate the salient features in the targeted layers in the DenseNet model architecture, we have introduced the application of Sp-RST and applied it to the extracted features. In this concern, we also aim to (3) investigate the effect of combining the classification outcome of classifiers trained on the Sp-RST selected hand-crafted features and the Sp-RST selected internal feature maps generated automatically inside the deep convolutional network. Accordingly, the capacity of this network for learning a combination of mammographic images is discussed.

This paper is structured as follows. Section II introduces the basic concepts of RST for feature selection. Section III details the application in breast cancer via the use of DenseNet and Sp-RST for large-scale data pre-processing. The experimental setup and the results are discussed in Section IV, and the conclusion is presented in Section V.

II. Rough Set Theory

Rough Set Theory (RST) [16] is seen as formal approximation of the conventional set theory that provides a filter based approach. This approach can extract knowledge from a problem domain in a concise way and retain the information content while reducing the involved amount of data.

A. Preliminaries of Rough Set Theory

In rough set theory, the training data set is called an *information table* that can be defined as a tuple $T = (U, A)$. U and A are two finite non-empty sets, where U refers to the *universe* of primitive instances (or objects) and A refers to the set of features. Each feature $a \in A$ is described with a set of values V_a named the *domain* of a. The feature set A can be partitioned into two sub-sets; namely the *conditional* feature set C and the *decision* attribute D. Let $P \subset A$ be a sub-set of attributes. The central concept to rough set theory is the *indiscernibility relation* which is denoted by $IND(P)$. $IND(P)$ is an equivalence relation that can be defined as follows: $IND(P) = \{(x, y) \in U \times U : \forall a \in P, a(x) = a(y)\}$, where $a(x)$ refers to the value of attribute a of the instance x. In case where $(x, y) \in IND(P)$ then x and y are said to be *indiscernible* with respect to the sub-set of features P. The set of all equivalence classes of $IND(P)$ is denoted by $U/IND(P)$, and refers to a partition of U which is determined by P. Every element in $U/IND(P)$ is a set of indiscernible instances with respect to the sub-set of features P. Based on these, the equivalence classes $U/IND(C)$ and $U/IND(D)$ can be defined and are named *condition* and *decision* classes, respectively. In RST, any $X \subseteq U$ and feature sub-set $R \subseteq A$, and using the knowledge of R, X can be approximated by two key concepts named the R-*lower* approximation and the R-*upper* approximation. The lower approximation of X describes the set of instances of U that are certainly in X. The R-*lower* approximation is defined as follows: $\underline{R}(X) = \bigcup\{E \in U/IND(R) : E \subseteq X\}$. On the other hand, the upper approximation of X describes the set of instances of U that are possibly in X. The R-*upper* approximation is defined as follows: $\overline{R}(X) = \bigcup\{E \in U/IND(R) : E \cap X \neq \emptyset\}$. The concept that defines the set of instances that are not certainly, but can possibly be classified in a specific way is named the *boundary region*. The latter is defined as follows: $BND_R(X) = \overline{R}(X) - \underline{R}(X)$. If $BND_R(X)$ is empty, meaning that $\overline{R}(X) = \underline{R}(X)$, then the X concept is said to be R-*definable*. In the opposite case, X is a *rough set* with respect to R. Another essential concept in RST is the *positive region*. The *positive region* of decision classes $U/IND(D)$ with respect to the set of the conditional features C is denoted by $POS_c(D)$; where $POS_c(D) = \bigcup \overline{R}(X)$. $POS_c(D)$ is a set of instances of U that can be classified with certitude to classes $U/IND(D)$ when using features of C. This means that $POS_c(D)$ describes the union of all the equivalence classes which are defined by $IND(P)$ that each, certainly, can specifically induce the decision class D. Based on the *positive region* concept, the *dependency of features* is defined as follows: $k = \gamma(C, c_i) = \frac{|POS_C(c_i)|}{|U|}$. The *dependency* measures the degree k of the dependency of a specific feature c_i on a set of features C.

B. Reduction Process

The theory of rough sets aims at finding the smallest sub-set of the conditional attribute set in a way that the resulting reduced database remains consistent with respect to the decision attribute. To achieve this, the theory defines the $Reduct$ concept and the $Core$ concept. A sub-set $R \subseteq C$ is said to be a D-*reduct* of C in the case where $\gamma(C, R) = \gamma(C)$ and there is no $R' \subset R$ such that $\gamma(C, R^{'}) = \gamma(C, R)$. Based on this formula, the $Reduct$ can be defined as the minimal set of selected features that preserve the same dependency degree as the whole set of features. In practice, from the given information table, it is possible that the theory generates a set of reducts: $RED_D^F(C)$. In this situation, any reduct in $RED_D^F(C)$ can be selected to describe the original information table. The theory also defines the $Core$ concept which is the set of features that are enclosed in all reducts. The $Core$ concept is defined as $CORE_D(C) = \bigcap RED_D(C)$, where $RED_D(C)$ is the D-reduct of C. More precisely, the $Core$ is defined as the set of features that cannot be omitted from the information table without inducing a collapse of the equivalence class structure. This means that all the features which are in the $Core$ are indispensable.

III. Application

A. Data set

Of critical concern for supervised learning, specially in deep learning approaches, is the amount of annotated data with labels for training the network. Currently, access to a large mammography repository that provides images with similar acquisition methods is not realistic, because providing such database is time-consuming, tedious and costly. Currently, four mammographic databases have become publicly available and these data repositories were used in this research to conduct experiments. The first and the second data sets are from the wide-ranging annotated Breast Cancer Digital Repository (BCDR) [19], containing digitised film (F03) and full field digital mammography images (D01) from women in northern Portugal. The third data set is a sub-set of the Digital Database for Screening Mammography (DDSM) [20] provided by the University of South Florida. The fourth data set that contains images acquired at a Breast Centre in Portugal is the Inbreast [21] repository providing full field digital mammography images. In this study, we concentrated on biopsy-proven mammographic mass lesions. Detailed information about these benchmarking data sets is provided in Table I.

1) Patch Extraction: To keep the information in the images acquired from various centres and decrease the sensitivity of classification models to intensity variations, a pre-processing approach was implemented. Firstly, all images were segmented into background and tissue (using a thresholding approach for digital images and the approach developed by Chen and Zwiggelaar for digitised images [22]). Subsequently, the intensity values of the segmented tissue regions were normalised. Using the provided annotations of identified lesions (manual contours annotated by clinical experts), a Region of Interest (RoI) was extracted with the size equal to double the square bounding box of the abnormality. The reason for this RoI selection was that not only the mass abnormality itself, but also its neighbourhood contained relevant information, which were considered by radiologists for diagnosis and has been reported to result in significant improvement in the final classification performance [11]. The prepared data sets were randomly split into training, validation and test sets as 65%, 25% and 10% of the whole database, respectively based on cases, ensuring that there was no women overlap between the splits. The distribution of patient characteristics in each data repository is provided in Table I.

2) Data Augmentation: Data augmentation was performed to alleviate the relatively small amount of training data for the deep learning approach. To achieve this, five random rotations were done. In order to keep the original morphology of the abnormality and avoid shape changes due to common re-sizing methods, square bounding boxes for the abnormalities were

TABLE I: Publicly available databases containing masses, used in this study. (MLO: mediolateral-oblique view; CC: Cranial-Caudal view)

	BCDR-F03 [19]	**BCDR-D01** [19]	**DDSM** [20]	**Inbreast** [21]
Number of cases	341	51	975	102
Number of images	664	105	1930	102
Benign images	369	69	1023	34
Malignant images	295	36	907	68
Resolution (bits/pixel)	8	14	12, 16	14
Image mode	digitised	digital	digitised	digital
View	MLO, CC	MLO, CC	MLO, CC	MLO, CC
Age distribution	58.4±15.3	57.7±13.5	58.9±11.5	-

considered instead of the abnormality bounding box, whilst extracting patches. The patches were scaled to 256×256. Then, random 224×224 crops followed by random mirroring were performed to generate more training samples. However, there was variation in the number of mammograms per case and not all images necessarily contained annotated abnormalities. As seen in Table I, there is an imbalance between the number of benign and malignant cases/images. Doing augmentation led to a further imbalance between these samples. To address this issue and improve the regularisation of the training procedure, random noise from one of the Gaussian, Localvar, Poisson, Salt & Pepper and speckle distributions was generated and added to the selected image in the training data set.

B. Deep Learning Methodology

1) DenseNet Architecture: CNNs have become the dominant type of models for image classification. Among the well-known CNNs, we have focused on DenseNet [13]. This was based on a comparative study conducted in [23], in which various types of deep networks (DenseNet, GoogLeNet, VggNet-16 and AlexNet) were compared with regards to the generalisation ability of the model to various data sets for the current problem. DenseNet is an interesting model because it uses the key characteristic of bypass signals from the preceding layers to the subsequent ones to enforce optimal information flow in the form of feature maps. This is done by concatenating features while disregarding redundant feature maps during training. Among the DenseNet variants [13], DenseNet-BC is a successful model proposed for the ImageNet [24] classification challenge. Since our images have the similar size as the ones fed into the DenseNet-BC structure, we have therefore selected it in our proposed approach, and we shortly name it "DenseNet" in this paper. This network is made up of L layers and each layer implements a specific non-linear transformation, which can be a composite function of different commonly used operations in deep learning concept such as Batch Normalisation, rectified linear units, Pooling and Convolution [6], [13]. In this model, direct connections from any layer to all subsequent layers are incorporated to enable the l^{th} layer to receive the feature-maps of all preceding layers. To facilitate down-sampling as an essential part of a convolutional networks, the network is divided into multiple densely connected blocks (dense-blocks), which are connected to each other through transition layers (composed of a batch normalisation layer, a 1×1 convolutional layer and a 2×2 average pooling layer). DenseNet's growth rate (k) is a new parameter of the network defined for generating narrower layers and is set to 4 to specifically refer to the DenseNet-BC structure (i.e. 4 dense-blocks and 3 transition layers). The initial convolution layer incorporates $2k$ convolutions of size 7×7 and the number of feature-maps in all other layers follow the setting for k. Each dense-block consists of different repetition of a sequence of sub-layers, i.e. dense-blocks 1, 2, 3 and 4 have 6, 12, 24 and 16 sub-layer sequences, respectively as shown in Figure 1. Each layer takes all preceding feature-maps as input. The final Softmax classifier makes a decision based on the created features in the network. The rest of the model's parameters with regards to the kernel, stride and padding sizes were kept as default as detailed in [13].

2) Training DenseNet: The objective of training in deep learning is to minimise the difference error between the network prediction and the expected output (defined by expert radiologists). This error is then flowed backwards through the network using the back-propagation procedure [6] leading the network parameter values to be updated. In our experiments and with respect to [13], the DenseNet model was trained via a stochastic gradient descent solver with the parameters set to Gamma = 0.1, momentum = 0.9 and weight-decay = 10^{-5}. We trained the model using mini-batches of size 8 (according to our hardware specifications) and an initial learning rate of 0.001 with 33% step down policy for 30 epochs. In our implementations, the ImageNet data was used to do the initial training of DenseNet, whilst the network was fine-tuned using a combination of all the data sets presented in Table I.

DenseNet layers are very narrow and a small set of feature maps are added to the collective knowledge of the network during training, while the rest of the feature maps are preserved unchanged. After training the network, the low-level, mid-level and high-level features were extracted from the last pooling layer of four main dense-blocks referred to as F-DB-1, F-DB-2, F-DB-3 and F-DB-4 (see Figure 1). These features had the following dimensionality (size): 118 800 (4.3GB), 65 536 (2.3GB), 42 849 (1.4GB) and 10 000 (420MB) for dense-block 1 to 4, respectively. To deal with this amount of data, a distributed version of rough set theory, named Sp-RST [15], was used for feature selection (details are given in Section III-E). Sp-RST is based on a parallel programming design that allows to tackle big data sets over a cluster of machines independently

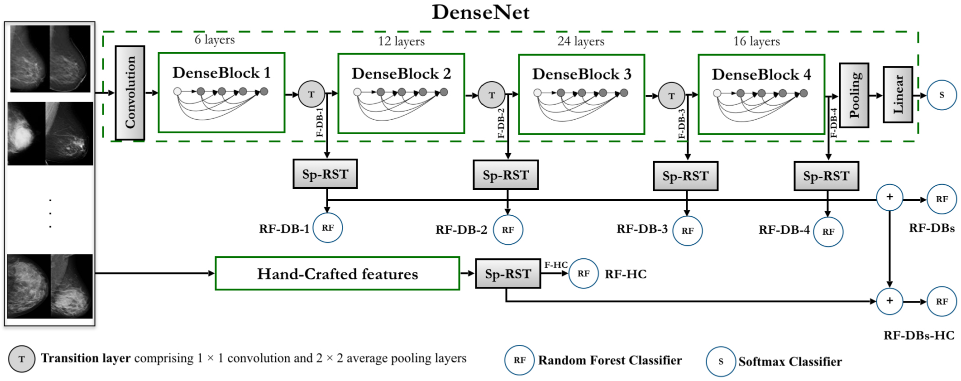


Fig. 1: Flowchart of the overall methodology.

from the underlying hardware/software. This feature selector indicated the most significant features from the input database, which corresponded to the low-level, mid-level and high-level deep features from different depths of the network. The application of Sp-RST decreased the number of features to 87 564 (3.7GB), 47 886 (2.1GB), 29 141 (1.3GB), and 7 326 (328.5MB) for dense-block 1 to 4, respectively.

C. Hand-Crafted Methodology

Adding to the previous process, several manual/hand-crafted features were extracted (see Figure 1 for F-HC) to be combined with the already selected deep learned features (i.e F-DB-1, F-DB-2, F-DB-3 and F-DB-4) to investigate the influence of these manual features on the overall classification performance of the model. The computed manual features are explained in Table II, where their selection was based on a comparative study made in [2]. A total of 2 570 (57.9MB) manual features were computed for each ROI and Sp-RST was also applied to decrease the number of manual features and could select 785 (19.1MB) significant features.

D. Training the Classifier

Using the selected hierarchical features from the network, four Random Forest (RF) classifiers (RF-DB-1, RF-DB-2, RF-DB-3 and RF-DB-4) were trained. The main motivation behind the use of RF was its robustness and efficiency in handling large data sets and its capability to not overfit the model. In our experiments, an RF parameter tuning step was performed to select the most adequate parameter values and these were set as follows: n.estimators = 500, max.features = 32 and max.depth = 6. The rest of the RF parameters were kept as default using the Sklearn Library[1]. A similar RF classifier was trained on hand-crafted features (RF-HC) to distinguish benign/malignant abnormalities. These five RF classifiers were trained to be used later in an ensemble method and the outcome of the respective RF classifiers were considered to give the final judgment, i.e. via the majority voting approach to choose the best model for the current problem. This was achieved while incorporating the effect of various features in the binary classification task of mass abnormalities. This will be further discussed in Section IV.

E. Sp-RST for Feature Selection

To select the most significant features from such amount of data, specifically from each input database referring to F-DB-1, F-DB-2, F-DB-3, F-DB-4 and F-HC, which corresponded to the feature pools extracted from each dense-block in DenseNet as well as from the hand-crafted feature pool—mapped as information data sets with respect to the rough set terminology—Sp-RST [15] proceeds as follows:

1) Problem Formalisation: Technically, the information table was first stored in an associated Distributed File System (DFS) which is reachable from any machine (computer) of the used cluster. To work on the input DFS in a distributed and parallel way, a Resilient Distributed Dataset (RDD) was created. We may formalise the latter as a given information table defined as T_{RDD}, where universe $U = \{x_1, \dots, x_N\}$ is the set of data items reflecting the extracted mammographic patches from the mammograms provided in Table I and as previously explained in Section III-A1, the conditional attribute set $C = \{c_1, \dots, c_V\}$ contains every single feature of the T_{RDD} information table and reflects pixel intensities from the feature maps for DenseNet corresponding to different internal filter responses, while for manual features it represents numerical values computed for each criteria covered in Table II. The decision feature D of our learning problem refers to the label (class) of each T_{RDD} sample having binary values d: either the patch is benign or malignant. D is defined as: $D = \{0, 1\}$. C presents the conditional attribute pool from where the most significant attributes were selected.

2) Feature selection process: For feature selection, the given T_{RDD} information table was partitioned first into m data blocks based on splits from the conditional attribute set C. Hence, $T_{RDD} = \bigcup_{i=1}^{m} (C_r) T_{RDD_{(i)}}$, where $r \in \{1, \dots, V\}$. Every $T_{RDD_{(i)}}$ was build using r random attributes which were selected from C, where $\forall T_{RDD_{(i)}} : \nexists \{c_r\} = \bigcap_{i=1}^{m} T_{RDD_{(i)}}$. Within a distributed implementation design, Sp-RST was applied to every $T_{RDD_{(i)}}$ while gathering

[1] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

TABLE II: Manual/Hand-crafted features extracted from each abnormality.

Manual feature	Description
Patient Based Features	Age and normalised area of the abnormality for each patient.
First Order Statistical Features	Mean, Standard Deviation, Variance, Skewness, Kurtosis of the abnormality for each patient. These features estimate the properties of individual pixels in the image without considering the spatial interaction between the image pixels.
Local Binary Pattern (LBP) Features	Represented by a histogram of each generated LBP code. LBP operator is a function of pair (P, R), where P tells the neighbourhood size and R is the radius. In the "Uniform" LBP, the pair values are set to (1,8) with 254 bins for the histogram.
Histogram of Oriented Gradient (HOG) Features	Counting occurrences of oriented gradient orientation in the image. For this, the magnitude and direction of the horizontal/vertical gradients are calculated in 32×32 divided blocks of image and the final HOG descriptor is represented by combining histograms (with 9-bins) from all the blocks.
Second order Statistical Features based on Grey Level Co-occurrence Matrix (GLCM)	Considers the relationship between the neighbouring pixels. In the current work, GLCM matrix for four directions (0, 45, 90, 145) and five pixel distances (0,1,2,3,4) were calculated. Afterwards Correlation, Energy, Contrast, Entropy, Variance and Homogeneity features were computed and the mean of these texture features from all matrices were added as final features

all the intermediate results from the distinct m created partitions. Technically, Sp-RST stars first of all by computing the indiscernibility relation for the decision class. We define the indiscernibility relation as $IND(D)$: $IND(d_i)$. Sp-RST will calculate $IND(D)$ for each decision class d_i by associating the same T_{RDD} data items (instances) that are expressed in the universe $U = \{x_1, \ldots, x_N\}$ and that belong to the same decision class d_i. This process is totally independent from the m created partitions. This is because the result depends on the class of the data instances, and not on the attribute set. Once this is calculated, Sp-RST builds the m random $T_{RDD_{(i)}}$ partitions as previously described. After that and within a specific partition, the algorithm generates first all the possible combinations of the C_r set of attributes, then calculates the indiscernibility relation $IND(AllComb_{(C_r)})$ for every created combination, and finally computes the dependency degrees of each attribute combination defined as $\gamma(C_r, AllComb_{(C_r)})$. Once all the dependencies are calculated, Sp-RST looks for the maximum value of the dependency among all the computed $\gamma(C_r, AllComb_{(C_r)})$. Let us recall that based on the RST preliminaries (seen in Section II), the maximum dependency refers to not only the dependency of the whole attribute set (C_r) describing the T_{RDD_i} but also to the dependency of all the possible attribute combinations satisfying the following constraint: $\gamma(C_r, AllComb_{(C_r)}) = \gamma(C_r)$. The maximum dependency reflects the baseline value for the feature selection task. In a next step, Sp-RST performs a filtering process to only keep the set of all combinations which have the same dependency degrees as the already selected dependency baseline value. In fact, through these computations, the algorithm removes in each level the unnecessary attributes that may negatively influence the performance of any learning algorithm. At a final stage, Sp-RST performs a second filtering process to only keep the set of combinations that have the minimum number of attributes. This is achieved by satisfying the full reduct constraints highlighted in Section II: $\gamma(C_r, AllComb_{(C_r)}) = \gamma(C_r)$ while there is no $AllComb^{'}_{(C_r)} \subset AllComb_{(C_r)}$ such that $\gamma(C_r, AllComb^{'}_{(C_r)}) = \gamma(C_r, AllComb_{(C_r)})$. Every combination that satisfies this constraint is evaluated as a possible minimum reduct set. The features defining the reduct set describe all concepts in the initial T_{RDD_i} training data set.

At the end of all these computations, the output of each created partition can be either only one reduct $RED_{i_{(D)}}(C_r)$ or a set (a family) of reducts $RED^F_{i_{(D)}}(C_r)$. As previously highlighted in Section II, any reduct among the $RED^F_{i_{(D)}}(C_r)$ reducts can be selected to describe the $T_{RDD_{(i)}}$ information table. Therefore, in case where Sp-RST generates a single reduct for a specific $T_{RDD_{(i)}}$ partition then the final output of this attribute selection phase is the set of features defined in $RED_{i_{(D)}}(C_r)$. These attributes represent the most informative features among the C_r features, and generate a new reduced $T_{RDD_{(i)}}$ defined as: $T_{RDD_{(i)}}(RED)$. The latter reduced base guarantees nearly the same data quality as its corresponding $T_{RDD_{(i)}}(C_r)$ which is based on the full attribute set C_r. In the other case where Sp-RST generates multiple reducts then the algorithm performs a random selection of a single reduct among the generated family of reducts $RED^F_{i_{(D)}}(C_r)$ to describe the corresponding $T_{RDD_{(i)}}$. This random selection is supported by the RST fundamentals and is explained by the same level of importance of all the reducts defined in $RED^F_{i_{(D)}}(C_r)$. More precisely, any reduct included in the family of reducts $RED^F_{i_{(D)}}(C_r)$ can be selected to replace the $T_{RDD_{(i)}}$ (C_r) attributes. At this level, the output of every i data block is $RED_{i_{(D)}}(C_r)$ which refers to the selected set of features. Nevertheless, since every $T_{RDD_{(i)}}$ is described using distinct attributes and with respect to $T_{RDD} = \bigcup_{i=1}^{m} (C_r) T_{RDD_{(i)}}$, a union operator on the generated selected attributes is needed to represent the original T_{RDD}. This is defined as $Reduct_m = \bigcup_{i=1}^{m} RED_{i_{(D)}}(C_r)$.

To further guarantee the Sp-RST feature selection performance while avoiding any critical information loss, to evolve the algorithm and to refine it, Sp-RST was run over N iterations on the T_{RDD} m data blocks and hence an output of N $Reduct_m$ is generated. Finally, an intersection operator applied on all the obtained $Reduct_m$ was required. This is defined as $Reduct = \bigcap_{n=1}^{N} Reduct_m$. Sp-RST could diminish the dimensionality of the original data set from $T_{RDD}(C)$ to $T_{RDD}(Reduct)$ by removing irrelevant and redundant features at each computation level. Sp-RST could also simplify the learned model and speed up the overall learning process. We invite the reader to refer to [15] for further details about the Sp-RST pseudo-code as well as details of its distributed tasks.

F. System Specifications

All deep learning based implementations were performed within the Caffe framework, using a NVIDIA GeForce GTX 1080 GPU on Intel Core i7-4790 Processor within Ubuntu 16.04. For feature selection and training the RFs, we performed the experiments on the High Performance Computing Wales using dual 12 core Intel Westmere Xeon X5650 and 36GB of memory to test Sp-RST, which was implemented in Scala 2.11 within Spark 2.1.1. We performed experiments for 652 partitions on 4 nodes and 10 iterations.

IV. Experimental Setup, Results and Discussion

The main aim of our experiments was to analyse the classification performance along with the behaviour of DenseNet. This was done using Sp-RST for the investigation of deep and hand-crafted features for mammography classification. Focusing on classification, the network performance was compared to several well-known state-of-the-art networks. This was to show DenseNet's effectiveness, generalisability and robustness toward different data repositories as will be discussed in Section IV-A. In the later stage, we have investigated the DenseNet behavior through a study of its signal passing through the targeted dense-blocks. This was done by comparing and discussing classification performances of each dense-block, hand-crafted features and various combinations on such features in ensemble techniques. Besides, we have introduced the use of t-SNE, for displaying the distribution of training and testing samples and the binary separation of classes. With the effective use of Sp-RST for feature selection, and by performing inference using activation maximization via regularized optimization [25], feature maps were visualised to give a better intuition of the DenseNet information flow. This will be discussed in Section IV-B. For the DenseNet evaluation performance, we have used the standard measures including accuracy (ACC), precision, recall, F1-score, and the Area Under Curve (AUC).

A. Classification Performance

We have empirically demonstrated DenseNet's effectiveness for binary classification of mass abnormalities on several mammographic data sets from multiple centers and vendors and with different protocols and compared with a set of well-known state-of-the-art architectures namely GoogLeNet [26], Vgg-Net [27] and AlexNet [5] in Table III. To ensure a fair comparison between these architectures, some factors such as differences in databases and data pre-processing were eliminated. Detailed explanations about the implementation and the models' parameters can be found in [23]. Based on our reported results presented in Table III, DenseNet outperformed the other mentioned models for the classification task in terms of ACC and AUC with 76% and 78%, respectively on the mixture of data sets (described in Table I). Taking this into account, the trained DenseNet was evaluated on testing samples of each database separately as well, as stated in Section III-A. Based on our experiments and the quantitative results provided in Table IV, DenseNet had the representational capacity to learn different mammographic abnormalities. Results also demonstrated the robustness and generalisability of this network to various image types (i.e., digitised and digital images as described in Table I) for the task of classification. Compared to other networks, DenseNet's superior classification performance could be associated to the innovative idea of short connections in the network structure. These connections enabled adding to the collective knowledge of the network during training while preserving the rest of the feature maps unchanged.

TABLE III: Comparing the classification performance for four well-known deep networks in terms of accuracy and area under curve. Results are obtained on the mixture of testing samples from all databases.

Evaluation	DenseNet	GoogLeNet	VggNet-16	AlexNet
ACC	**0.76**	0.72	0.75	0.72
AUC	**0.78**	0.72	0.77	0.67

TABLE IV: Classification of DenseNet for each mammographic data set.

Data set	Accuracy (%)	AUC
DDSM	73.50	0.82
BCDR-F03	84.48	0.78
BCDR-D01	100.00	1.00
Inbreast	81.82	0.75

B. Network Behaviour

As shown in Figure 1, different RFs were trained by feeding them various features from the hierarchical dense-blocks, e.g., F-DB-1 is fed to RF-DB-1, F-DB-2 is fed to RF-DB-2, etc., and from the manual features, i.e. F-HC is fed to RF-HC. As previously explained in Section III-E, these features were selected by Sp-RST. After this feature selection task and after training the RFs classifiers as highlighted in Section III-D, testing samples were fed to these classifiers. For analysis purposes, the independent classification performance of various dense-blocks (RF-DB-1, RF-DB-2, RF-DB-3, and RF-DB-4) and manual features (RF-HC) are presented in Table V.

Comparing the first four RF classifiers in Table V, results show that through the different layers the classification performance is gradually improving from the initial dense-block (RF-DB-1 with 52.3%) to the final dense-block (RF-DB-4 with 72.8%); where the difference is increased by nearly 20% in terms of classification accuracy. We notice the same behavior for the precision, recall, and F1-score with 16%, 21% and 32% improvements, respectively. Such interesting behaviour can be explained by the existing input concatenation in DenseNet architecture, which enables the feature maps learned in the preceding layers to be accessed by all the subsequent layers. This characteristic has encouraged feature reuse throughout the network.

To gain a better intuition about this feature reusing and hence the network behaviour, Figure 2 is given which shows the low-level (dense-blocks 1 & 2), mid-level (dense-block 3) and high-level features (dense-block 4) for a sample (patch) abnormality from the test set that were extracted from the final pooling layer of each dense-block. From Figure 2, in the low-level feature maps, as expected, projection to pixel space and feature map has revealed responses corresponding to edges and boundaries. Bypassing these feature maps to the subsequent dense-block and reusing them, the number of less informative or less discriminative responses was decreased and eventually in the final layers it can be seen that an optimal information flow in the form of feature maps was achieved. This can be explained by the fact that all of the feature maps were counted as important and salient features based on our Sp-RST feature selection approach. From an Sp-RST perspective, these selected features correspond to the reduct set and they represent the minimal set of selected attributes that preserve the same dependency degree as the whole set of attributes.

Figure 3 shows a visualisation of the Sp-RST feature distribution performed by t-SNE to gain a further intuition about the network's behaviour for the training and testing samples. From Figure 3, we can similarly see that the initial dense-blocks have generated filter responses during training containing a few salient and discriminative features, not leading to a separative boundary. This boundary reflects the lowest results in terms of classification performance (mainly in RF-DB-1 and RF-DB-2 with 51.8% and 52.7% in terms of classification accuracy) as presented in Table V. Besides, based on the results given in Table V, the best classification performance of a combination of testing samples was achieved in the final layer of DenseNet (RF-DB-4 with 72.8% in terms of classification accuracy), presenting a clear separative boundary as can be clearly seen in Figure 3. This is explained by the fact that each layer takes all preceding feature-maps as input, which allows features to be re-used throughout the network during training. Consequently, the model is able to learn more compactly and accurately.

Considering precision and recall values, we observed that the model in each layer (RF-DB-1 to RF-DB-4) is coping better with benign samples than malignant ones and the number of false negative predictions compared to the false positive predictions is higher for the network performance in each experiment but still improving as it goes deeper. Eventually, these two rates become the same for RF-DB-4 and RF-DBs but then it is affected by the subsequent blocks (global pooling and softmax (S)) in the standard DenseNet model, resulting in more false negatives than false positives (80% and 71% in term of precision and recall, respectively).

In order to investigate the effect of signal bypassing throughout the model, we have combined the Sp-RST salient opted features from the four dense-blocks (comprising low-level to high-level) and have fed such features to a similar RF classifier (RF-DBs) as seen in Figure 1. Possibly the most noticeable trend is to compare the classification results obtained by RF-DBs and the classification results obtained by RF-DB-4, which demonstrates the optimal information flow due to feature reusing and keeping maximum correlation independency. This is confirmed by the results shown in Table V, where we can notice that RF-DBs achieves a classification performance of 72.7% which is very similar to the classification performance given by the last dense-block, RF-DB-4 (72.8%). Moreover, at the end of the fourth dense block, a global average pooling was used following a softmax classifier (S) to represent the standard DenseNet. On the other hand, in our implementation, as ensemble during testing, if the binary classification outcome of low-level to high-level classifiers (RF-DB-1 to RF-DB-4) and the Softmax output (S) were the same, then that judgment was taken. Otherwise, the image was counted as an uncertain case and the output of the RF classifier trained on the selected manual features (RF-HC) was taken into account and the majority voting approach was applied, i.e. among the 6 classifiers (RF-DB-1 to RF-DB-4, S, and RF-HC). Via this approach, we aimed to evaluate if by adding the hand-crafted features to the model's intermediate abstracts, the classification performance of DenseNet could be improved.

Based on our experimental results and from Tables IV and V, we noticed that the classification performance using deep convolutional networks is significantly superior compared to using merely hand-crafted features. Due to low classification performance of RF-HC (in terms of accuracy, precision, recall and F1-score), and combing that in ensemble methods (RF-DBs-HC) has led to the model deviation and thus low results were obtained as stated in Table IV. Table V shows that in terms of classification accuracy the ensemble approach (71.2%) did not further improve the performance over the original DenseNet model (S) (76.5%). This was due to the low classification performance of classifiers trained on merely low-level features (RF-DB-1 and/or RF-DB-2 with accuracies equal to 51.8% and 52.7%, respectively) or manual features (RF-HC with 52.3%) compared to the network itself with 76.5%. Meanwhile, it is important to recall that the use of Sp-RST as a feature selection technique was mainly to understand the DenseNet behaviour and thus a comparison of RF-DBs to the softmax classifier (S) was not realistic.

Based on the conducted experiments, we highlighted the effectiveness of the use of Sp-RST as a feature selection technique when dealing with large data sets. Sp-RST could perform optimal feature selection while being able to analyse the facts hidden in data and find a minimal knowledge representation without requiring any additional information about the given data. Another important interest of our proposed solution was the use of t-SNE for large data visualisation. Both of these techniques helped to better understand and investigate the optimal information flow during the DenseNet learning process. These innovative aspects have not been done before and are the key contributions for developing a more suitable mammographic CAD system.

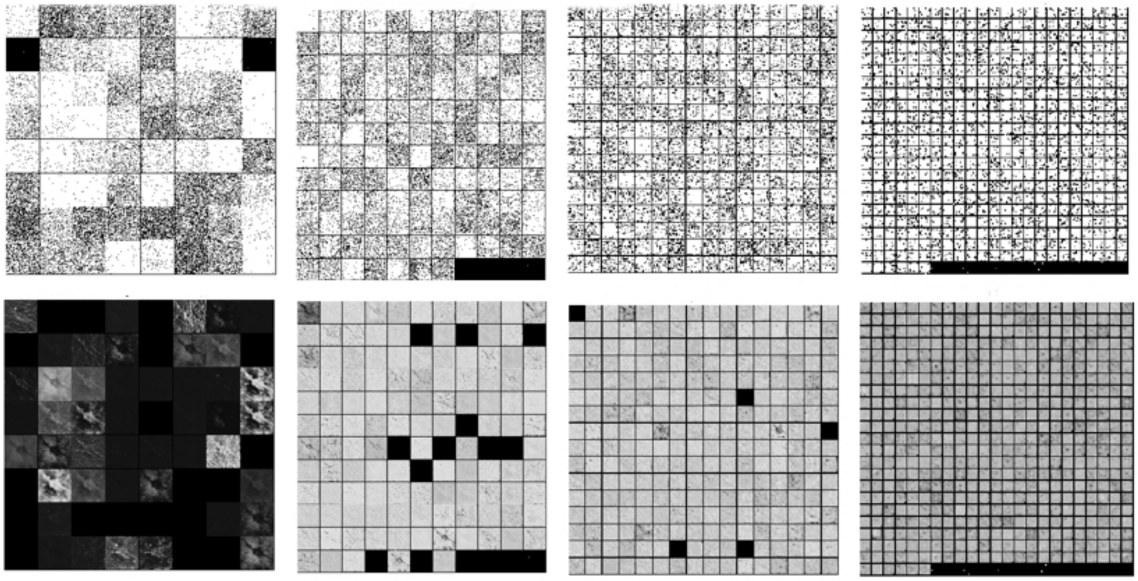


Fig. 2: Features extracted from the final pooling layer of each dense-block. The Sp-RST feature selection approach was applied to the extracted deep features. The first row shows the significant pixel-wise features. The second row is the result of opted features-maps. From left to right: F-DB-1, F-DB-2, F-DB-3 ans F-DB-4 in DenseNet.

TABLE V: Classification results based on various types of extracted features on the mixture of testing samples.

	Accuracy	Precision	Recall	F1-Score
RF-DB-1	0.523	0.57	0.52	0.41
RF-DB-2	0.527	0.57	0.52	0.41
RF-DB-3	0.655	0.71	0.66	0.63
RF-DB-4	0.728	0.73	0.73	0.73
RF-DBs	0.727	0.72	0.72	0.72
RF-HC	0.60	0.65	0.60	0.58
RF-DBs-HC	0.68	0.67	0.68	0.68
Softmax (S)	0.767	0.80	0.71	0.76
Ensemble approach	0.712	0.70	0.71	0.70

V. Conclusion

In this work, a pipeline for mammography mass classification was investigated by fine-tuning the DenseNet on a combination of various mammographic data sets, acquired from different vendors. Therefore, the capacity of this network for learning a combination of these databases was compared with the existing state-of-the-art models. On the other hand, we trained several random forest classifiers separately and in ensemble techniques. In these studied scenarios, using the distributed rough set based feature selection approach for mainly understanding the DenseNet behaviour, different levels of features extracted from the trained network and hand crafted features were opted and fed to the classifiers. Using proper visualisation techniques (i.e. sample distributions and feature maps), the insight into the function of intermediate feature layers and the operation of a dense network in flowing optimal and salient information was discussed. We concluded that end-to-end learning with DenseNet showed the representational capacity to learn the class of mammographic abnormalities and the mid-level to high-level features or manual features could not further improve the classification performance. However, larger and more diverse but coherent data sets are required to learn a more generalised model. Considering that in this dense convolutional structure, each layer receives supervision from the loss function through the shorter connections, attaching classifiers (i.e. RF) to every internal hidden dense-block can enforce the mid-level layers to learn more discriminative features, which can be suggested as future work.

Acknowledgment

Part of the research has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 702527. The authors would like to gratefully acknowledge Dr. Alun Jones and Sandy Spence for their support and maintenance of the GPU and the systems used for this research. Additional thanks go to the Supercomputing Wales project, which is part-funded by the European Regional Development Fund via the Welsh Government.

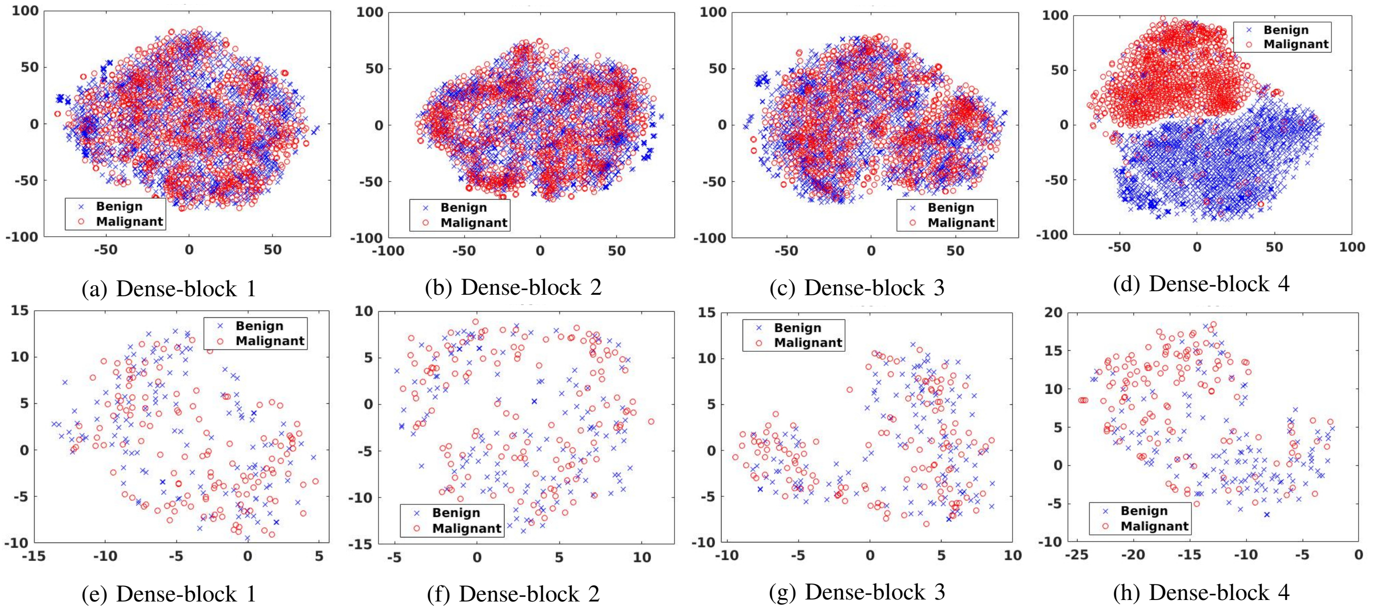
References

[1] B. W. Stewart and P. Kleihues, *World Cancer Report*. Lyon, France: IARCPress, International Agency for Research on Cancer, WHO, 2014.

[2] W. He, A. Juette, E. R. Denton, A. Oliver, R. Martí, and R. Zwiggelaar, "A review on automatic mammographic density and parenchymal segmentation," *International Journal of Breast Cancer*, vol. 2015, p. Article ID: 276217, 2014.

[3] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, and R. Zwiggelaar, "Deep learning in mammography and breast histology, an overview and future trends," *Medical Image Analysis*, vol. 47, pp. 45–67, 2018.

[4] A. Oliver, J. Freixenet, J. Marti, E. Perez, J. Pont, E. R. Denton, and R. Zwiggelaar, "A review of automatic mass detection and segmentation in mammographic images," *Medical Image Analysis*, vol. 14, no. 2, pp. 87–110, 2010.



(a) Dense-block 1 (b) Dense-block 2 (c) Dense-block 3 (d) Dense-block 4

(e) Dense-block 1 (f) Dense-block 2 (g) Dense-block 3 (h) Dense-block 4

Fig. 3: The scatter plot distribution of sample images in training (row 1) and testing (row 2), by t-SNE using features extracted from different dense-blocks.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[7] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in *IEEE International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2015, pp. 1–8.

[8] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9351. Springer, 2015, pp. 652–660.

[9] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *Journal of Medical Imaging*, vol. 3, no. 3, p. 034501, 2016.

[10] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," *Medical Image Analysis*, vol. 35, pp. 303–312, 2017.

[11] A. Hamidinekoo, Z. Suhail, T. Qaiser, and R. Zwiggelaar, "Investigating the effect of various augmentations on the input data fed to a convolutional neural network for the task of mammographic mass classification," in *Annual Conference on Medical Image Understanding and Analysis*. Springer, 2017, pp. 398–409.

[12] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 248–257, 2016.

[13] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[14] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Springer Science & Business Media, 2012, vol. 454.

[15] Z. C. Dagdia, C. Zarges, G. Beck, and M. Lebbah, "A distributed rough set theory based algorithm for an efficient big data pre-processing under the spark framework," in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017, pp. 911–916.

[16] L. Polkowski, *Rough sets in knowledge discovery 2: applications, case studies and software systems*. Physica, 2013, vol. 19.

[17] L. Polkowski, S. Tsumoto, and T. Y. Lin, *Rough set methods and applications: new developments in knowledge discovery in information systems*. Physica, 2012, vol. 56.

[18] Z. C. Dagdia, C. Zarges, B. Schannes, M. Micalef, L. Galiana, B. Rolland, O. de Fresnoye, and M. Benchoufi, "Rough set theory as a data mining technique: A case study in epidemiology and cancer incidence prediction," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), Proceedings Part III, LNAI Vol. 11053 (to appear)*. Springer, 2018.

[19] M. G. Lopez, N. Posada, D. C. Moura, R. R. Pollán, J. M. F. Valiente, C. S. Ortega, M. Solar, G. Diaz-Herrero, I. Ramos, J. Loureiro, T. C. Fernandes, and B. M. Ferreira de Arajo, "BCDR: a breast cancer digital repository," in *15th International Conference on Experimental Mechanics*, 2012.

[20] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," in *In Proceedings of the 5th International Workshop on Digital Mammography*. Medical Physics Publishing, 2001, pp. 212–218.

[21] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[22] Z. Chen and R. Zwiggelaar, "A combined method for automatic identification of the breast boundary in mammograms," in *5th International Conference on Biomedical Engineering and Informatics (BMEI)*. IEEE, 2012, pp. 121–125.

[23] A. Hamidinekoo, Z. Suhail, E. Denton, and R. Zwiggelaar, "Comparing the performance of various deep networks for binary classification of breast tumours," in *Workshop on Breast Imaging (IWBI 2018)*, vol. 1071807, 2018, p. 6.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[25] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.