Applying Deep Learning for the Detection of Abnormalities in Mammograms

Steven Wessels and Dustin van der Haar

Abstract Medical imaging produces massive amounts of data. Computer aided diagnosis (CAD) systems that use traditional machine learning algorithms to derive insights from the data provided in the medical industry struggle to perform at a competent level regarding sensitivity and false positive minimization. This paper looks at some of the current methods used to improve CAD systems in the domain of forming breast cancer diagnosis with mammograms. This paper presents deep learning models that use Convolutional Neural Networks (CNN) to identify abnormalities in mammographic studies that can be used as a tool for the diagnosis of breast cancer. We run two experimental cases on two public mammogram databases, namely MIAS and the DDSM. Firstly, the abnormality severity was classified. Secondly, the combination of abnormality type and its severity were compared in multi-label classification. Two CNN architectures, namely miniature versions of VGGNet and GoogLeNet, were also compared. We were able to achieve a best AUC of 0.85 for the classification of abnormality severity on the DDSM data set and a best Hamming loss of 0.27 on the MIAS data set for the multi-label classification task.

Keywords Deep learning · Convolutional neural networks · Medical imaging

1 Introduction

It was estimated that in 2017, the health care industry generated 150 exabytes of data and that by 2020, that figure will increase to 2300 exabytes [1]. Researchers at IBM postulated that medical imaging constitutes over 90% of medical data [2]. A primary source of image generation in health care are mammograms, a conventional

S. Wessels (✉) · D. van der Haar
Academy of Computer Science and Software Engineering,
University of Johannesburg, Johannesburg, Gauteng, South Africa
e-mail: swessels@jhb.dvt.co.za

D. van der Haar
e-mail: dvanderhaar@uj.ac.za

K. J. Kim and H.-Y. Kim (eds.), *Information Science and Applications*,
Lecture Notes in Electrical Engineering 621,
https://doi.org/10.1007/978-981-15-1465-4_21

means of screening breast cancer. Traditionally, mammograms would have to be inspected by a radiologist for signs of breast cancer. Manual inspection is an error-prone, costly, and time exhausting task. To alleviate the challenges associated with manual inspection, computer aided detection and diagnosis systems that used pattern recognition and learning algorithms for inspection were designed and deployed [3]. By 2008, a reported 74% of all mammography examination were screened using CAD [4]. The effectiveness of early CAD systems was usually compromised by the lack of discriminative power of the classifiers that were used and the high computational cost of performing the inspection task. Additionally, CAD systems were used in problem specific domains and were biased towards how the system programmer believed the interpretation task ought to be performed, and as a result, CAD systems produced high false positive rates. Recently, deep learning models have been used to boost accuracy and to alleviate the previously mentioned problems.

The rest of this paper is structured in the following manner: Sect. 2 highlights the severity and prevalence of breast cancer currently and discusses problems with current CAD systems. This section will also briefly discuss modern approaches to improving the performance of CAD systems. Section 3 is a literature review surrounding the applications of machine learning in CAD systems for mammographic analysis. The data sets used to evaluate the model are discussed in Sect. 4. Section 5 provides an in depth description of the models presented in this paper. Section 6 presents the results obtained by the presented deep learning models, as well as a discussion surrounding these results. The paper concludes with Sect. 7, where a summary of the article is given along with a brief review of how this work could be taken forward.

2 Problem Background

2.1 Mammography and Breast Cancer

Breast cancer is the most prevalent cancer among the female population worldwide. Mammography is used to detect breast cancer and has shown to reduce mortality due to breast cancer by 38–48% [5]. A mammogram examination captures images of each breast from two angles using a low-intensity x-ray. These images are inspected for lesions, both malignant and benign, characteristic masses, and microcalcifications [6]. Zonder and Smithuis outlined the standard reporting procedure for mammographic screening [7]. Firstly, the indication of the screening must be detailed. Secondly, breast composition is described. Thirdly, any significant findings, which could be masses, asymmetry, calcifications, or distortions, need to be specified. Then a comparison against previous screenings can be made, and an assessment can be made by assigning a Breast Imaging Reporting and Data System (BI-RADS) category. The most prominent problem that exists with a diagnosis from mammograms, and indeed is pervasive throughout cancer imaging, is the human error inherent in radiology. The possible source of errors may be fatigue, distraction, inexperience, or an insufficient

number of previous cases to make a correct diagnosis. The authors of [8] estimated that up to 30% lesions could be missed during screening. These factors led to the emergence of CAD systems.

2.2 Deep Learning

The techniques applied for identifying breast cancer in early CAD systems made use of support vector machines (SVM), k-nearest neighbour (KNN), and linear discriminant analysis (LDA) [6]. State-of-the-art techniques make use of deep learning, a new method from the domain of artificial intelligence and machine learning. Deep learning uses neural networks that are constructed with multiple hidden layers to improve and enhance the recognition accuracy of various data types, particularly images in the case of convolutional neural networks (CNN). Due to the deep architecture of these networks, representations that would have been previously hidden can be discovered and fundamentally enhance classification accuracy [9]. A large amount of data is required to train a deep learning model, causing its widespread use on large data sets. Deep convolutional neural networks have yielded excellent results in medical applications, as is shown in the related work section.

3 Related Work

The use of deep learning methodologies on mammograms in the identification of breast cancer is currently a popular research area, and a fair amount of research on applying deep learning in this problem domain has been published in recent years. In 2016, Kooi et al. performed a comparative study between traditional CAD systems and CNN's at high and low sensitivity [3]. Their data set consisted of 45,000 images. The authors never specified whether this data set was publicly available but we assume that it was privately collected by their institution. The effects of different preprocessing methods were compared, such as augmentation and manual segmentation. They found that data augmentation slightly improved CNN performance. Additionally, CNN's slightly outperformed the traditional CAD system at low sensitivity but had comparable performance at high sensitivity. Ultimately, the best performing CNN achieved an area-under-curve (AUC) score of 0.941. Later in 2016, Wang et al. researched identifying microcalcifications and other lesions in mammograms to improve the diagnostic accuracy of all microcalcifications [6]. The authors used a semi-automated segmentation technique to categorize all calcification types and a discrimination classifier model to assess accuracy. The authors stated their models were tested on a large data set consisting of over 1200 samples generated from mammograms collected between 2011 and 2015 at SunYat-sen University Cancer Center (Guangzhou, China) and Nanhai Affiliated Hospital of Southern Medical University (Foshan, China). An accuracy of 87.3% was achieved,

1.5% better than the best performing SVM model. Jain and Lèvy produced the final major work of 2016 that was associated with deep learning in mammography. To our knowledge, they were the first authors to attain state-of-the-art results with a CNN on a publicly available data set, namely the Digital Database for Screening Mammography (DDSM). The DDSM is the largest publicly available data set of mammograms to our knowledge [10]. They evaluated and compared two prominent CNN architectures, AlexNet and GoogLeNet, to a shallow baseline CNN. Overall, the GoogLeNet architecture achieved the best accuracy score of 92.9%, slightly outperforming AlexNet, which achieved 89.0% accuracy, while the baseline network could only achieve 60.4% accuracy [11]. In 2017, Mohamed et al. aimed to use deep learning to distinguish between scattered and heterogeneous density masses found in mammograms [9]. The model used was CNN based in conjunction with an extensive mammogram data set. The data set that was used was of 22,000 mammogram images collected from their institution. The metric that the authors of [9] used to evaluate their system was an AUC score from a receiver operator (ROC) curve. A final AUC score of 0.9882 was achieved. In 2018, Kim et al. wanted to assess the feasibility of using a data-driven imaging bio-marker (DIB), which features a deep CNN algorithm, for use in mammography [4]. Again, the data set used was based on mammograms collected at a private institution. The authors were able to prove the viability for DIB in mammography by achieving AUC scores of 0.903 and 0.906 for the test and validation sets, respectively. Finally, a recent work that utilized a Faster R-CNN was proposed by Ribili et al. They trained and tested their models on public data sets. The model was trained on the DDSM and validated with the INbreast data set. They achieved an AUC of 0.95 with the validation set. Although some of these systems achieve remarkable results, many of them were trained and tested on private data sets, giving a little context to their respective models. Furthermore, in all the cases above, binary classification was used, whether it was classifying abnormalities as malignant or benign, or classifying the class of the abnormality. Our paper will use only publicly available mammogram data sets and will investigate the performance of multiple CNN architectures in single and multiple label scenarios to make a fair comparison among methods.

4 Experimental Setup

Medical data is inherently highly sensitive, despite this, there exist several publicly available data sets for mammography. To evaluate our models, we used the mammographic image analysis society (MIAS) digital mammogram database and the Digital Database for Screening Mammography (DDSM). MIAS is the oldest available mammogram data set and has been widely used in literature to assess a variety of approaches. MIAS consists of 322 mediolateral oblique images in a portable gray map (.pgm) format [12]. Additionally, MIAS provides a label for the severity of a mammographic study, and the class label of any abnormalities present. We will take advantage of these labels to implement a multi-label classifier. The MIAS data set

is split into 208 normal, 63 benign, and 51 malignant cases. The DDSM is a much larger data set than MIAS and will be the primary data set used to evaluate the models presented in this paper. The DDSM consists of 2620, four view (mediolateral oblique and craniocaudals of each breast), mammographic studies which amounts to 10,480 images in total [10]. The mammographic studies found in the DDSM were collated from four separate hospitals in the USA, with different digitizers being used to convert the the film-based studies to a lossless JPEG format. The images found in the DDSM may contain chain code overlays and each digitizer introduces it's own noise and artifacts to an image. Overlayed ground truth information regarding suspicious regions in an image were provided by both data sets. However, we have chosen only to consider the overall classification of a mammographic exam (i.e. normal, benign, or malignant) during training, as to remove as much domain knowledge from the model as possible.

5 Model

5.1 Pre-processing

Both the .pgm and ljpeg formats of the MIAS and DDSM data sets respectively are incompatible with many image processing and deep learning APIs and need to be converted to a compatible format. All images were converted into .png (portable network graphics) format while maintaining the original resolution. These .png images are resized to smaller dimensions, in our case 128×128 pixels due to RAM limitations, and flattened before being loaded into memory. The raw pixel intensities of the images are scaled to the range [1], and the labels associated with the images are then also loaded into memory. Once all necessary data is in memory, the data set can be split into sections for training and testing. We used a 70–30 split for testing and training.

5.2 Convolutional Neural Networks

This paper looks to evaluate the performance of two neural network architectures and their multi-label variants. The neural network architectures that we investigated were scaled down architectures of the VGGNet and GoogLeNet, the implementation details of which can be found in [13]. These architectures are visualized in Fig. 2a, b respectively.

The mini VGGNet features the ReLU (Rectified Linear Unit) activation function and also uses batch normalization, max pooling and dropout. Batch normalization can be quite effective at minimizing the number of epochs required to train the neural network and works by normalizing a given inputs activations before that input moves to the networks next layer [14]. The pooling layers reduce the spatial dimensions of

the input volume. Dropout is used to minimize overfitting by randomly disconnecting neurons between layers. We used 25% dropout for both architectures. The final layer of VGGNet has a softmax activation classifier.

The mini GoogLeNet makes use of convolution modules and inception modules. The convolutional modules are responsible for applying a convolutional filter, after which batch normalization and activation take place. At the inception module, branching occurs. The branching is because the convolutions are applied in parallel, and the features derived are concatenated. An example of a small inception module can be seen in Fig. 1. Downsampling is used to reduce spatial dimensions. Just like the mini VGGNet, a softmax classifier is used in the last layer.

During training, the parameters within each network are optimized using stochastic gradient descent (SGD), and the batch size was set to 32. The initial learning rate was set to 0.005. Categorical cross-entropy was used as the loss function. For the case of multi-label classification, the final layer used a sigmoid activation function for classification (Fig. 2).

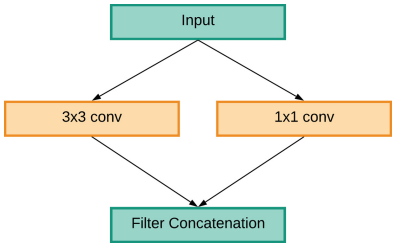

Fig. 1 Visualization of an inception module for the MiniGoogLeNet architecture

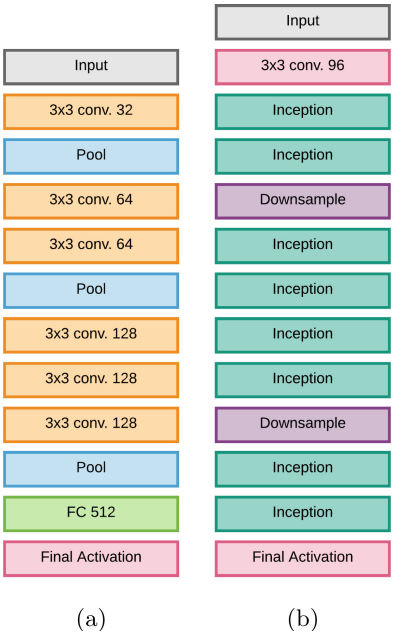

Fig. 2 **a** The mini VGGNet architecture. **b** The mini GoogLeNet architecture. The final activation layer is a softmax activation for a single label classification class or a sigmoid activation for a multi-label classification task

6 Results

To generate results with these models, we would run the models under two experimental cases. Firstly, the models would perform a multi-class classification task over the DDSM data set. The class labels would correspond to the types of mammographic studies found in the data set, namely normal, benign, and malignant cases. To measure loss across training and validation, categorical cross-entropy was used. Other metrics collected to evaluate model performance included precision and recall. Finally, the multi-label architectures would be applied to the MIAS data set so that the type of abnormality could be classified, along with its severity. A sigmoid activation layer replaced the softmax activation layer used in previous experiment to achieve multi-label classification. Binary cross-entropy was used to evaluate loss during training and validation of the models, and a Hamming loss metric was calculated for each model. All experimental cases were run over 30 epochs, with the same hyperparameters mentioned in Sect. 5.

6.1 CNN Classification Comparison for Abnormality Detection

In [15], the authors compared the performance of common CNN architectures on mammographic data set in cases where models had been trained from-scratch and pre-trained models. One of the data sets they used was the CBIS-DDSM, a well curated subset of the original DDSM. Although the architectures used in [15] are the deeper variants of ours, it is still useful to use their results for the VGG-16 and GoogLeNet architectures as a reference point for our results.

The results for the first experimental case are summarized in Table 1 and the confusion matrices for each model's performance can be seen in Fig. 4. Additionally, we have included each networks ROC curve in Fig. 3. In comparison to the trained from-scratch VGG-16 network used in [15], our mini VGGNet performed slight poorer with a best AUC of 0.68 in comparison to their AUC of 0.702. The training accuracy's were very close, with our model achieving 59.47% to their 58%. From Fig. 4a, we can see that the mini VGGNet struggled to discriminate between benign and malignant cases. The mini GoogLeNet also struggled, although to a lesser de-

Table 1 Performance comparison of the presented architectures on the DDSM data set

Architecture	Accuracy		Loss		Recall	Precision (%)
	Test (%)	Val (%)	Test	Val		
miniVGGNet	59.47	46.12	0.84	1.51	48.23%	46.19
miniGoogLeNet	76.19	66.16	0.55	0.72	66.84%	67.63

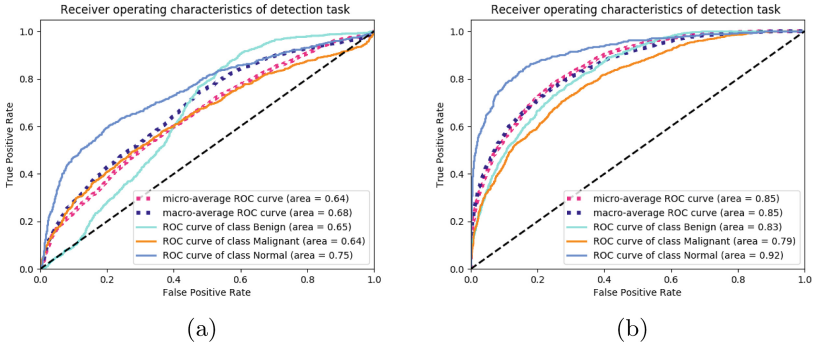

Fig. 3 Comparison of the ROC curves generated by each network on the DDSM where **a** is the mini VGGNet and **b** is the mini GoogLeNet

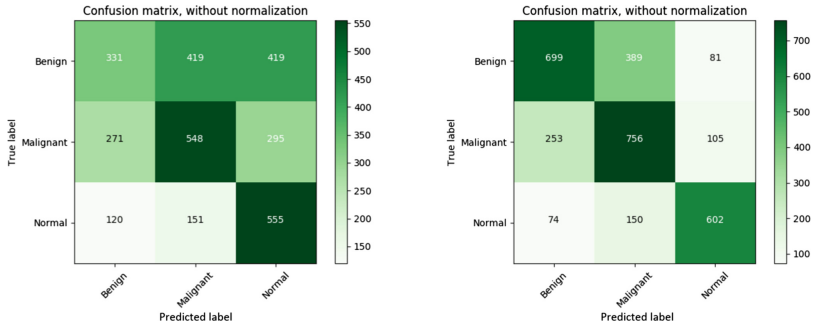

Fig. 4 Comparison of the confusion matrices generated by each network on the DDSM

gree, to distinguish between benign and malignant cases. However, in comparison to the GoogLeNet used in [15], the mini GoogLeNet demonstrated generally good performance with an AUC score of 0.85 and training accuracy of 76.19%. Additionally, the validation precision and recall rates were satisfactory. The GoogLeNet of [15] attained an AUC score of 0.59 and an accuracy of 59.8%, although it is worth noting that they used a slower learning rate and only trained their network for 12 epochs. Regarding the difference in performance between the two models we have evaluated here, we suspect that the difference may be due, in part, to the use of the same dropout percentage on both networks. Larger networks usually benefit more from the use of dropout, and because the VGGNet is a shallower network than GoogLeNet, it may have been wiser to use less regularization on the mini VGGNet [16].

Table 2 Performance comparison of the presented architectures on the MIAS data set

Architecture	Accuracy		Loss		Hamming loss
	Test (%)	Val (%)	Test	Val	
miniVGGNet	97.68	75.82	0.07	0.84	0.27
miniGoogLeNet	85.50	78.02	0.35	1.27	0.3

6.2 MIAS Multi-label Classification

For the final experiment, the MIAS data set was used because of the labels provided by the data set. The CNN architectures were modified to make a multi-label classification through the use of a final sigmoid activation layer. The results for each network are summarized in Table 2. Here, we see that the mini VGGNet outperformed the mini GoogLeNet. The accuracy achieved by the mini VGGNet, 97.68%, is very high and suggests that overfitting is taking place. The MIAS data set has a small number of samples, especially for deep learning tasks, and the small size of the data set may be causing the models, especially the mini VGGNet, to fail to generalize well. Moreover, MIAS is an imbalanced data set which undermines the use of the accuracy metric as the primary performance measure. A commonly used metric to evaluate multi-label classification is Hamming loss. Hamming loss computes a value between 0 and 1 that indicates how many times on average, the relevance of an example to a class label is incorrectly predicted [17]. The Hamming loss of each architecture does not indicate the same levels of performance as the accuracy scores do, but are still satisfactory nevertheless.

7 Conclusion and Future Work

In this paper, we presented two deep learning models that could be used to (a) classify abnormalities in mammographic studies and (b) provide a multi-label description of a mammogram regarding the severity and class of abnormalities. The mini VGGNet struggled to classify abnormalities in the DDSM data set and only achieved an AUC of 0.68. On the MIAS data set for multi-label classification, the mini VGGNet indicated signs of overfitting but did achieve an acceptable Hamming loss value. The mini GoogLeNet demonstrated satisfactory performance for both experimental cases. From-scratch training on limited medical data set for models to perform tasks such as multi-class and multi-label abnormality classification is a great challenge for neural networks. The miniature GoogLeNet we tested here demonstrated promising results, and taking this work forward, we may investigate from-scratch performance of other deep CNN architectures such as ResNet. Deep learning solutions to detect indicators of cancer in mammograms have achieved excellent results in the litera-

ture and could be extended to other biomedical imaging problems, like lung cancer screening, because deep learning provides versatile solutions that are applicable to a wide range of problem domains.

References

1. Health M (2018) Healthcare data generation in surge mode—Moxe health. (online) Moxe health. Available at: http://www.moxehealth.com/2017/07/19/expect-surge-healthcare-data-requests/. Accessed 16 Aug 2018
2. Ali A (2018) Deep learning applications in medical imaging (present use cases). (online) TechEmergence. Available at: https://www.techemergence.com/deep-learning-applications-in-medical-imaging/. Accessed 16 Aug 2018
3. Kooi T, Litjens G, van Ginneken B, Gubern-Mrida A, Snchez C, Mann R, den Heeten A, Karssemeijer N (2017) Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 35:303–312
4. Kim E, Kim H, Han K, Kang B, Sohn Y, Woo O, Lee C (2018) Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. Sci Rep 8(1)
5. Ribli D, Horvth A, Unger Z, Pollner P, Csabai I (2018) Detecting and classifying lesions in mammograms with Deep Learning. Sci Rep 8(1)
6. Wang J, Yang X, Cai H, Tan W, Jin C, Li L (2016) Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci Rep 6(1)
7. Zonderland H, Smithuis R (2013) Bi-RADS for mammography and ultrasound
8. Calas M, Gutfilen B, Pereira W (2012) CAD e mamografia: por que usar esta ferramenta? Radiol Bras 45(1):46–52
9. Mohamed A, Berg W, Peng H, Luo Y, Jankowitz R, Wu S (2017) A deep learning method for classifying mammographic breast density categories. Med Phys 45(1):314–321
10. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer P (2001) The digital database for screening mammography. In: Proceedings of the fifth international workshop on digital mammography, pp 212–218
11. Jain A, Lèvy D (2016) Breast mass classification from mammograms using deep convolutional neural networks. In: Conference on neural information processing systems (NIPS 2016), pp 1–6
12. Suckling J, Parker J, Dance D, Astley S, Hutt I, Boggis C, Ricketts I et al (2015) Mammographic image analysis society (MIAS) database v1.21 (Dataset). https://www.repository.cam.ac.uk/handle/1810/250394
13. Rosebrock A (2017) Deep learning for computer vision with python, 1st ed. PyImageSearch, pp 130–140
14. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML'15 proceedings of the 32nd international conference on international conference on machine learning, vol 37, pp 448–456
15. Tsochatzidis L, Costaridou L, Pratikakis I (2019) Deep learning for breast cancer diagnosis from mammograms a comparative study. J Imaging 5(3):37
16. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res (online) 15:1929–1958. Available at: http://jmlr.org/papers/v15/srivastava14a.html
17. Sorower MS (2010) A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis, pp 1–25