Computers in Biology and Medicine 107 (2019) 18–29

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

RAMS: Remote and automatic mammogram screening

Timothy Cogan*, Maribeth Cogan, Lakshman Tamil

Quality of Life Technology Laboratory, Department of Electrical and Computer Engineering, The University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX, 75080, USA

ARTICLE INFO

Keywords:
TensorFlow
Artificial neural network
Convolutional
Deep learning
Faster R-CNN
DDSM
INbreast
Mammograms
Telemedicine
SVM

ABSTRACT

About one in eight women in the U.S. will develop invasive breast cancer at some point in life. Breast cancer is the most common cancer found in women and if it is identified at an early stage by the use of mammograms, x-ray images of the breast, then the chances of successful treatment can be high. Typically, mammograms are screened by radiologists who determine whether a biopsy is necessary to ascertain the presence of cancer. Although historical screening methods have been effective, recent advances in computer vision and web technologies may be able to improve the accuracy, speed, cost, and accessibility of mammogram screenings. We propose a total screening solution comprised of three main components: a web service for uploading images and reviewing results, a machine learning algorithm for accepting or rejecting images as valid mammograms, and an artificial neural network for locating potential malignancies. Once an image is uploaded to our web service, an image acceptor determines whether or not the image is a mammogram. The image acceptor is primarily a one-class SVM built on features derived with a variational autoencoder. If an image is accepted as a mammogram, the malignancy identifier, a ResNet-101 Faster R-CNN, will locate tumors within the mammogram. On test data, the image acceptor had only 2 misclassifications out of 410 mammograms and 2 misclassifications out of 1,640 non-mammograms while the malignancy identifier achieved 0.951 AUROC when tested on BI-RADS 1, 5, and 6 images from the INbreast dataset.

1. Introduction

Worldwide, breast cancer is the most common cancer for women and the second most common cancer even when considering both men and women. In addition, breast cancer is the leading cause of cancer death in women across the globe [1]. For women as young as 40, mammograms, x-ray images of breast tissue, can provide a cost effective means for breast cancer screening [2]. Typically, mammograms are screened by radiologists who determine whether or not a biopsy is needed to classify a tissue abnormality as malignant or benign [3,4]. Machine learning algorithms capable of performing at or above the level of a radiologist could potentially replace or assist radiologists and thereby reduce the cost of screenings as well as lead to earlier and more reliable detection of breast cancer. For the past decade, many different machine learning algorithms have been proposed using state-of-the-art techniques to aid in the detection and classification of malignant abnormalities [5].

In mammogram analysis as in other areas of research, deep learning has been emerging as a dominant machine learning technique because it helps deal with the challenge of feature extraction. Machine learning techniques such as support vector machines are effective when there is a strong feature set, but in problems such as malignancy identification there is a fundamental challenge in determining the relevant features of an abnormality. Convolutional neural networks are particularly effective in extracting features from images and are being leveraged in a majority of deep learning solutions in mammogram research [5–7]. A few studies also leverage the faster region-based convolutional neural network (Faster R-CNN) architecture which is discussed later in this paper. For example, Ribli et al. achieved state-of-the-art performance on the INbreast dataset by use of a VGG-16 based Faster R-CNN [8]. [9,10] also present the effectiveness of VGG-based architectures. As an alternative to using a Faster R-CNN, [9] demonstrated an approach for adapting a patch classifier into a whole image classifier. This approach has an advantage in that the whole image classifier can be trained without annotations for each malignancy region of interest. A summary of recent approaches in applying deep learning to the problem of mammogram classification are listed in Table 2. The purpose of our work is to improve upon past research by proposing a complete, end-to-end solution for mammogram screenings. A primary component in our system is a web application which provides a simple and accessible interface for doctors and other health care professionals. Also critical to our solution is a machine learning algorithm which identifies whether

* Corresponding author.
E-mail addresses: tcc090020@utdallas.edu, timothy.cogan@utdallas.edu (T. Cogan), mxr127730@utdallas.edu (M. Cogan), laxman@utdallas.edu (L. Tamil).

https://doi.org/10.1016/j.compbiomed.2019.01.024
Received 11 December 2018; Received in revised form 25 January 2019; Accepted 25 January 2019
0010-4825/ © 2019 Elsevier Ltd. All rights reserved.

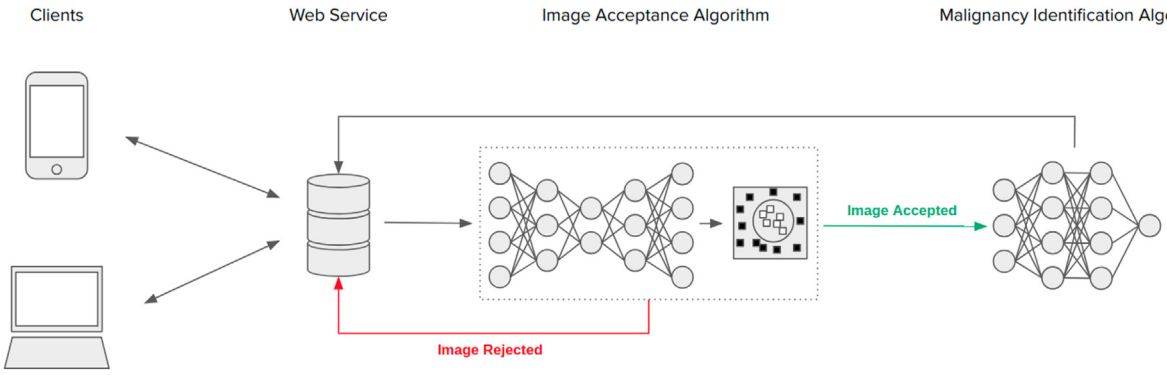

Fig. 1. System overview for the telemedical web application.

or not an image is truly a mammogram. Once an image has been submitted to the web application, this image acceptance algorithm filters images before they are passed along for further analysis. This further analysis is performed by a tumor identification network which locates all malignancies and feeds these locations to a generated report. A diagram depicting our total solution is shown in Fig. 1.

The image acceptor algorithm is relatively unique among existing literature, driven by the demands of a widely accessible web application. The authors are unaware of any similar algorithms for mammogram identification at the time of this writing. In contrast and as previously mentioned, much research has been performed in the area of malignancy identification. Our tumor identification network attempts to be an improvement upon related studies which also use a Faster R-CNN architecture by leveraging a ResNet-101 rather than VGG-based neural network [8,10]. We chose a ResNet-101 Faster R-CNN because it has been shown to provide a good trade-off between accuracy and training time [11]. Our approach is also unique in that the tumor identifier was trained and validated on mammograms that were enhanced with contrast limited adaptive histogram equalization (CLAHE) as well as color mapped from greyscale to RGB using a sequential color mapping scheme [12].

2. Telemedical web application

The web application provides an intuitive graphical user interface by which clients can submit mammograms for analysis, receive generated reports, and provide feedback about their experience. On the welcome page of the website, two buttons are presented to the user. One button allows users to select a local image for analysis while the other button submits a request for analysis. Fig. 2 shows the welcome page by which users are able to upload and submit an image for analysis. Following an image analysis, results are returned to the user as shown in Fig. 3. The result contains fields with the filename, number of abnormalities identified, overall probability of malignancy, malignancy locations, processing time, and date of analysis. Following a submission, the user is presented with an opportunity to provide feedback which is then stored in a database. The user interface was developed using HTML and JavaScript while the server side logic was developed with Python and the Flask web development framework. Python was a natural choice for developing the back end since the machine learning algorithms were also developed with Python.

3. Image acceptor

3.1. Image acceptor background

The image acceptor verifies that each submitted image actually depicts a mammogram. As an example for why this is important, if a user were to submit a picture of a cat, the malignancy identifier could erroneously identify the cat as a malignant abnormality. This is because the malignancy identification network has only been trained to perform accurately when presented with a mammogram. When presented with an image that is not a mammogram, the malignancy identifier has unpredictable behavior. Therefore, the image acceptor is necessary so that the malignancy identification network is never presented with a non-mammogram. Although identifying a cat as a tumor is not ideal, an even greater problem could arise if a non-mammogram medical image was submitted to the network. For example, if an image of a brain MRI was submitted to the network, an invalid classification could be misinterpreted by anyone who does not understand the scope of the malignancy identifer. Because our web application could be accessible to a wide audience, potentially anyone with internet access, the image acceptor is necessary to prevent malicious users from discrediting the system and to provide assurance that users will only use the system as intended.

Building an image acceptor is challenging because the variety of non-mammogram images is virtually limitless. Although we could assemble our own dataset of mammograms and non-mammograms and train a binary classifier to distinguish between the two classes, this approach might require a large and diverse set of non-mammograms to avoid overfitting the non-mammogram dataset. In addition, we may have low confidence that the network will perform well if presented with a class of images it has never seen before. As an extreme example, if the non-mammogram class consisted solely of cat images, the classifier would really be a mammogram versus cat classifier instead of a mammogram versus non-mammogram classifier. Ideally, we would like to teach an algorithm to recognize mammograms without the use of any non-mammogram images. The challenge of building such a network can be considered a one-class classification problem. Although recent studies have proposed novel strategies for leveraging deep learning in one-class problems [13,14], there is much opportunity for advancement in this segment of classification. The approach suggested in the following sections is relatively simple but also reasonably effective and very fast to train and evaluate.

3.2. Image acceptor methodology

The image acceptor algorithm is shown in Fig. 4. The two primary components of the network are the autoencoder and scorer. The autoencoder is the component which learns what a mammogram looks like whereas the scorer scores how closely a given image fits the model captured by the autoencoder. That is, the autoencoder has been trained for encoding and decoding mammograms and in that sense knows what a mammogram looks like, whereas the scorer doesn't know anything about mammograms but simply judges the quality of the encoding/decoding process. Although both of these components could be developed as neural networks, only the autoencoder was developed as a neural network for the system proposed in this paper. The autoencoder is a variational autoencoder and the scorer utilizes a one-class SVM which classifies each autoencoding as effective or ineffective. Due to

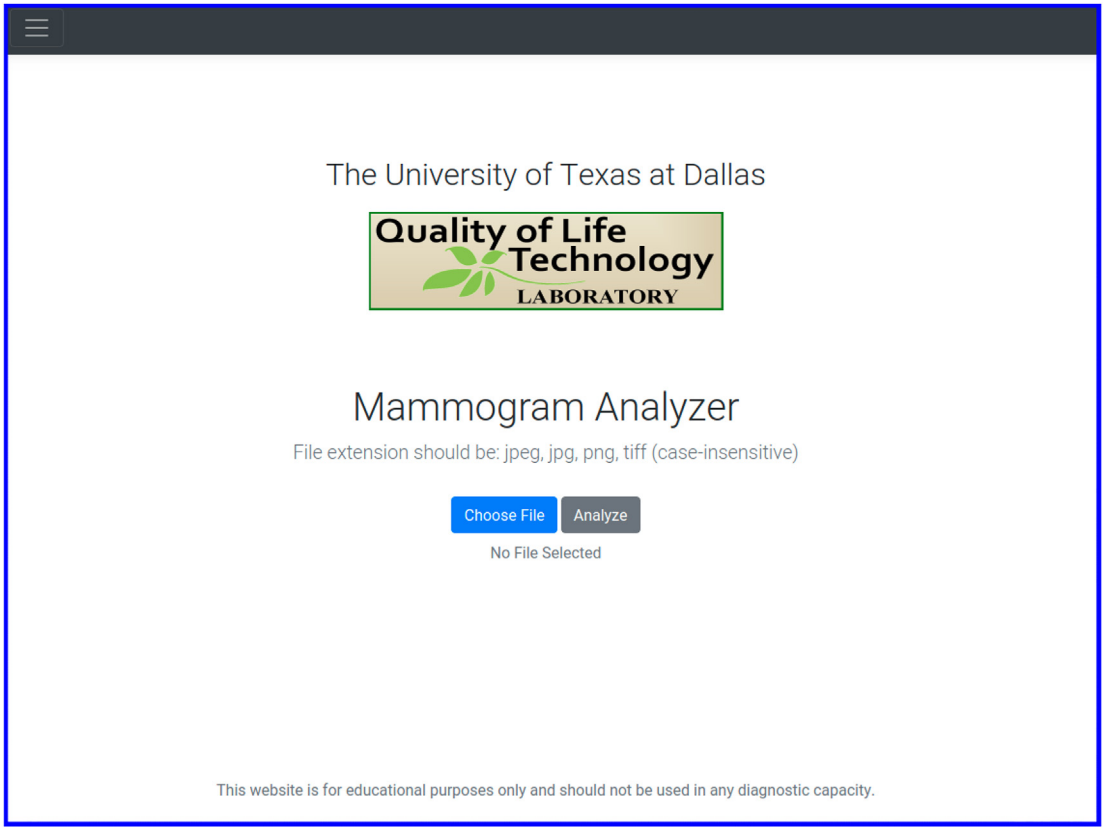

Fig. 2. Welcome page for the telemedical web application.

the nature of these components, only images from the DDSM were necessary in training the system and no non-mammogram images were necessary for training.

In more detail, the image acceptor can be represented in 5 steps: normalize the mammogram orientation, encode/decode the mammogram, threshold the original and decoded mammograms, calculate the MSE pixel difference between the original and decoded mammograms, and pass 1 minus the MSE values to a one-class SVM. The first step, normalizing mammogram orientation, reduces the burden on the autoencoder without affecting overall system performance. To normalize orientation, each input image is multiplied, pixel-by-pixel, by an image that is black on the left side and white on the right side and an image that is white on the left side and black on the right side. Next, the pixel-by-pixel multiplications are summed for the black-on-left and black-on-right image cases to produce two results. The input image is flipped if and only if the black-on-left, white-on-right multiplication produces the highest result. In the second step, the orientation-normalized image is first encoded, compressed, and then decoded, uncompressed, by a variational autoencoder. If the input image is a mammogram, the autoencoder will do a decent job encoding/decoding but if the input image is not a mammogram, the autoencoder will do a very poor job. More detail regarding the autoencoder will be discussed later in this paper. In the third step, the original and autoencoded images are thresholded. Thresholding is an unsupervised clustering technique by which lighter pixels are saturated to the highest possible value and darker pixels are floored to 0. The threshold is image dependent and is calculated as the average of cluster means from a 2-means clustering algorithm. In the fourth step, the original and autoencoded images are compared via pixel-by-pixel mean squared error (MSE). This MSE operation is performed for both thresholded and non-thresholded pairs of images and can be written as $1 - mean((x0 - x1)^2)$ where $x0$ and $x1$ represent pixel value arrays of the two images being compared. In the fifth and final step, one minus each of the two MSE values is passed to a one-class SVM which declares the autoencoding to be effective or ineffective and produces a final label of mammogram or not mammogram.

Out of these five steps, the variational autoencoder is perhaps the most complex. An autoencoder is a neural network that attempts to encode or compress an input into a minimal number of values and then decode or uncompress these values to reproduce the original input. A variational autoencoder is a type of autoencoder where the input is compressed into a set of mean and standard deviation values. These mean and standard deviation values are used to modify samples drawn from a normal distribution, and these modified samples are fed into the decoder to reconstruct the original input. For the variational autoencoder used in this study, a convolutional neural network layer followed by four fully connected layers is used to encode each input image into two mean and two standard deviation values. Four fully connected decoder layers are then used to recreate the input image. The encoder layers are progressively smaller while the decoder layers are progressively larger – exact layer sizes can be seen in Fig. 5 [15–17]. An additional benefit to using the variational autoencoder is that it can be repurposed as a mammogram generator. Generated mammograms could be useful for training other neural networks or could be used as a component in virtual patient generation for training medical students. Generated mammograms could be free from U.S. Health Insurance Portability and Accountability Act (HIPAA) regulations and sampled at any time. An example of the variational autoencoder being used as a mammogram generator is shown in Fig. 7.

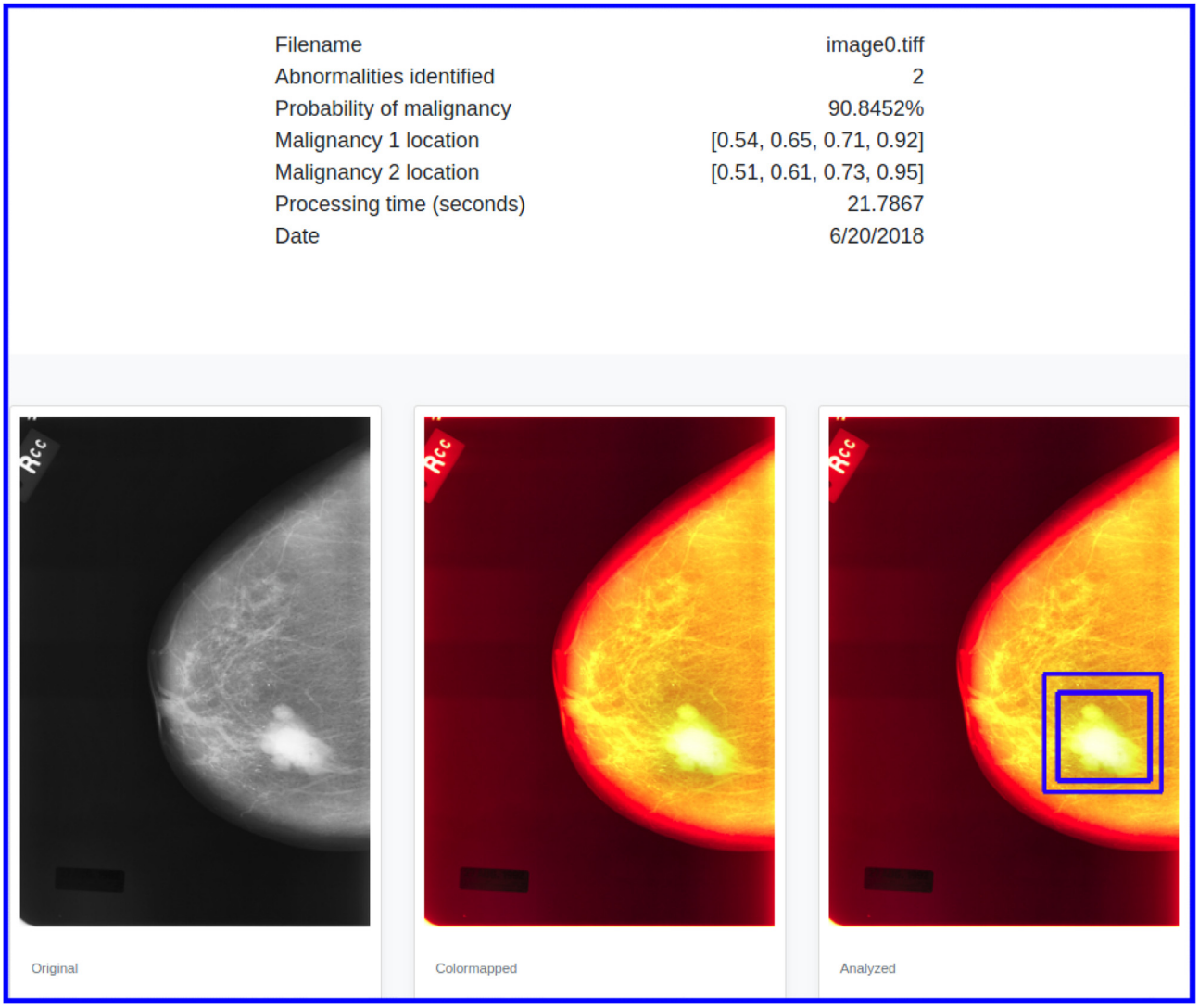

Fig. 3. Results page for the telemedical web application.

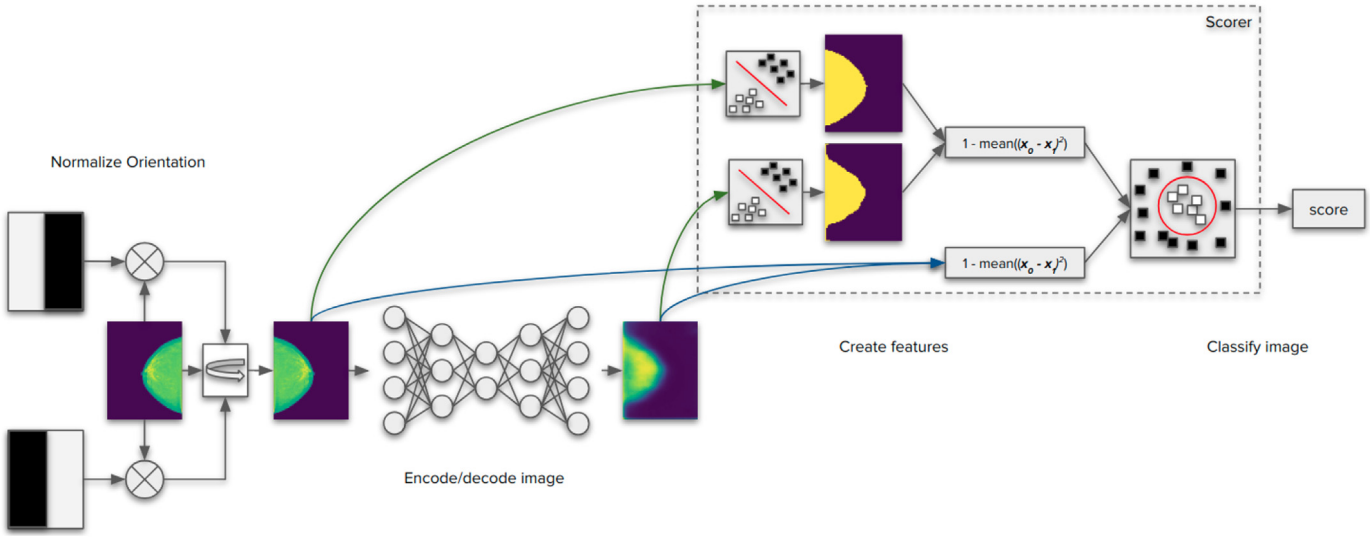

Fig. 4. Image acceptor architecture.

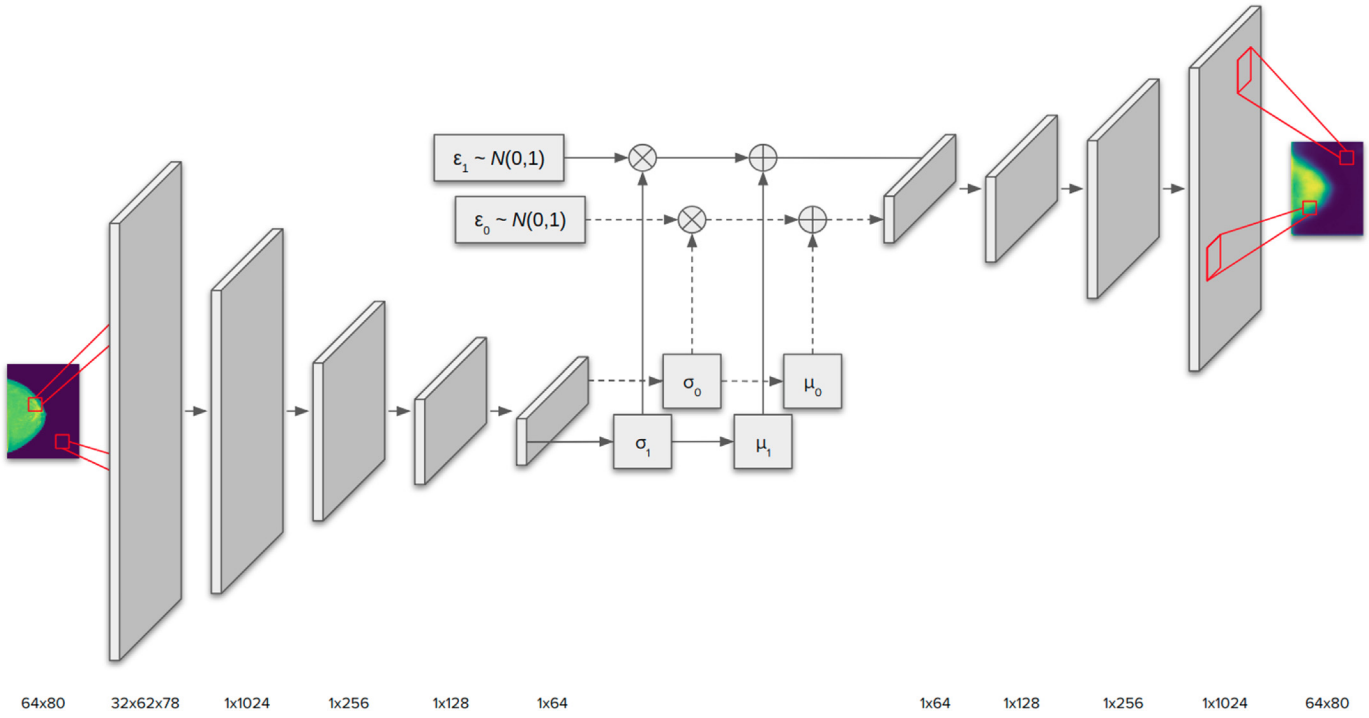

Fig. 5. Architecture of the image acceptor variational autoencoder.

Algorithm 1: Image Acceptor (Mammogram Classifier)

Input: Any jpg, png, dcm, or tiff image
If image is color, convert to greyscale
Resize image to 64x80
Normalize image orientation
Feed image through variational autoencoder
Threshold original and autoencoded images
for $pair$ *in* $[pair_{thresholded}, pair_{non-thresholded}]$ **do**
  Calculate $1 - mean((x_0 - x_1)^2)$
Feed $1 - mean((x_0 - x_1)^2)$ values to one-class SVM

3.3. Image acceptor results

A scatter plot with points corresponding to various image classes is shown in Fig. 6 where features 0 and 1 correspond to the mean squared pixel errors calculated according to Fig. 4. Five different classes of images were tested against the identification system where each class contained 410 images. The five classes were mammogram (taken from the INbreast dataset), gastrointestinal (Z-line images taken from the Kvasir dataset) [18], chest x-ray [19], brain MRI [20], and miscellaneous (random images taken from the COCO 2014 validation dataset) [21]. As can be seen in Fig. 6, the one-class SVM is able to distinguish most mammograms from non-mammograms. Only 2 out of 410 test mammograms and 23 out of 2,812 train mammograms were misclassified while only 2 out of 1,640 non-mammograms were misclassified. These results are also summarized as a confusion matrix in Table 1. The system presented here is efficient and effective, and there is also likely room for improvement. As stated before, the significance of this network is that it is trained solely on positive examples of mammograms. No examples of non-mammograms were used in the training of the system. Training and evaluation required 7 minutes 12 seconds and 4 minutes 7 seconds, respectively, on an Intel i7 processor.

3.4. Image acceptor discussion

As mentioned previously, a couple of recent studies by Chalapathy and Perera have approached the issue of one-class classification with state-of-the-art deep learning techniques [13]. Although the techniques and ideas from typical one-class classification problems provide good reference, the challenge of identifying mammograms is unique. For example, Perera built a model to identify abnormal chairs amid images of normal chairs. In Perera's study, the challenge comes from the fact that examples of abnormal chairs, the positive class, are small to non-existant and very diverse. A 2014 study by Ganesan et al. [22] actually investigates one-class classification for mammograms, but the proposed system is designed for benign versus malignant mammogram differentiation. As in Perera's study [14], a lack of training data is the primary challenge addressed. In the case of identifying mammograms, examples of the positive class, mammograms, are plentiful but examples of the negative class, every other type of image in the world, are too numerous and diverse to easily handle. Therefore, we believe the work we've done with mammogram identification, image acceptance, is a strong contribution to existing classification literature.

4. Tumor identifier

4.1. Tumor identifier methodology

The Digital Database for Screening Mammography (DDSM) and the INbreast database were used for training and validation, respectively. The INbreast database was very easy to work with; however, the DDSM contains many more images [23,24]. Training was performed solely on the DDSM while validation was performed on the INbreast dataset. Mammograms containing completely normal tissue or malignant abnormalities were used from these two data sets. All of the images used for either training or validation were mapped from greyscale to RGB space via a sequential color mapping scheme. The first reason for this mapping is that color mapping can enhance the visual dynamic range of an image. By utilizing the entire RGB space, images can convey greater visual information about relative intensities than can be done using

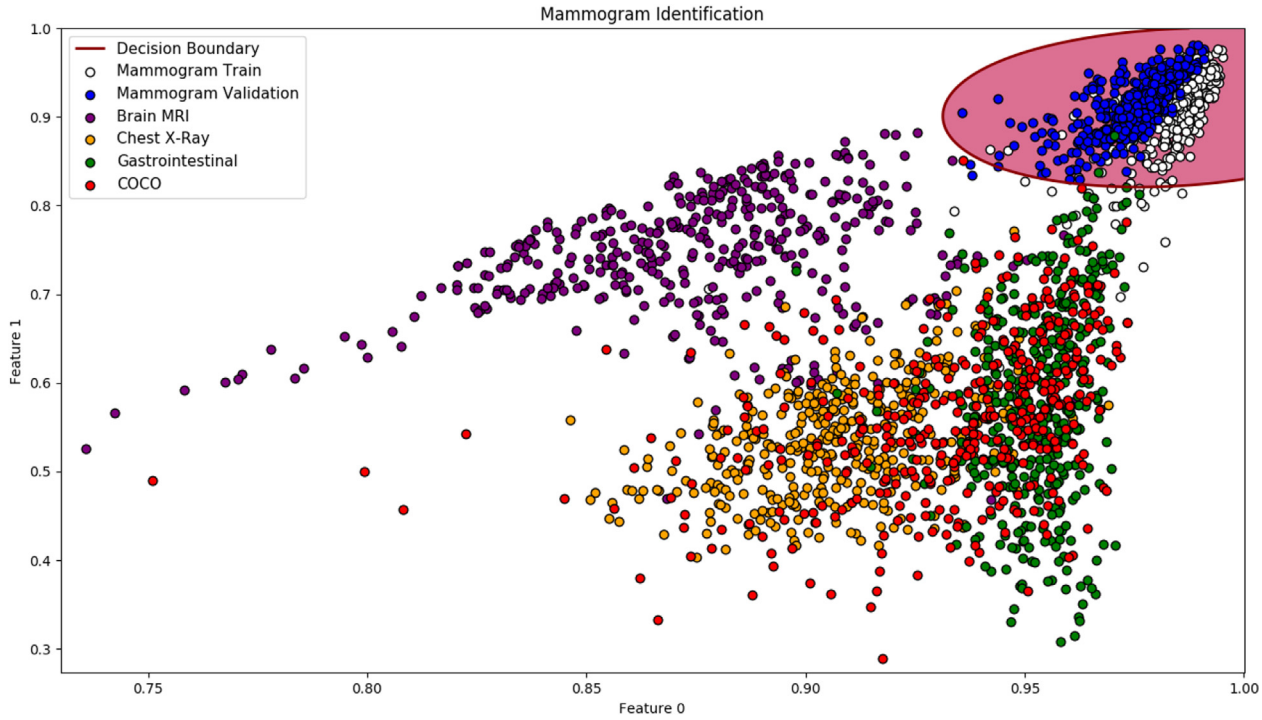

Fig. 6. Separation between 5 different image classes for the image acceptor.

greyscale. Because the neural network used in this paper was pre-trained on RGB images taken from the COCO dataset, we conjectured that the network would receive a similar benefit from color mapping that humans receive from color mapping. The second reason for this mapping is due to the fact that the mammograms were converted to 24-bit RGB JPEG images for input into the network. Mammograms may be up to 12-bit or 16-bit greyscale, and mapping from 16-bit greyscale to greyscale in 24-bit RGB space requires truncation of bits. That is, 24-bit RGB space can only represent 8-bit greyscale. By color mapping mammograms, relative pixel intensities can be represented by more than 8-bits in 24-bit RGB space. Therefore, more information regarding relative pixel values is maintained through color mapping. Although there are many different color map choices, a sequential color mapping scheme is used because it causes pixel lightness to increase with pixel value and is thereby visually intuitive. For example, low pixel values will map to a *dark* red while high pixel values will map to a *light* yellow [12]. Examples of color mapped mammograms can be seen in the results section of this paper.

In addition to colormapping, contrast limited adaptive histogram equalization (CLAHE) is used to improve mammogram quality by enhancing image contrast. CLAHE enhances image contrast by equalizing pixel value distributions. In other words, CLAHE will make dark pixels darker and light pixels lighter such that the dynamic pixel range is more fully leveraged. In the case of mammograms, CLAHE is able to enhance the visibility of edges and features critical for identifying tumors. Oftentimes, breast tissue pixels are clustered and potentially saturated in the highest range of possible pixel values. By equalizing the range of these tissue pixels, pertinent tissue features become easier to identify.

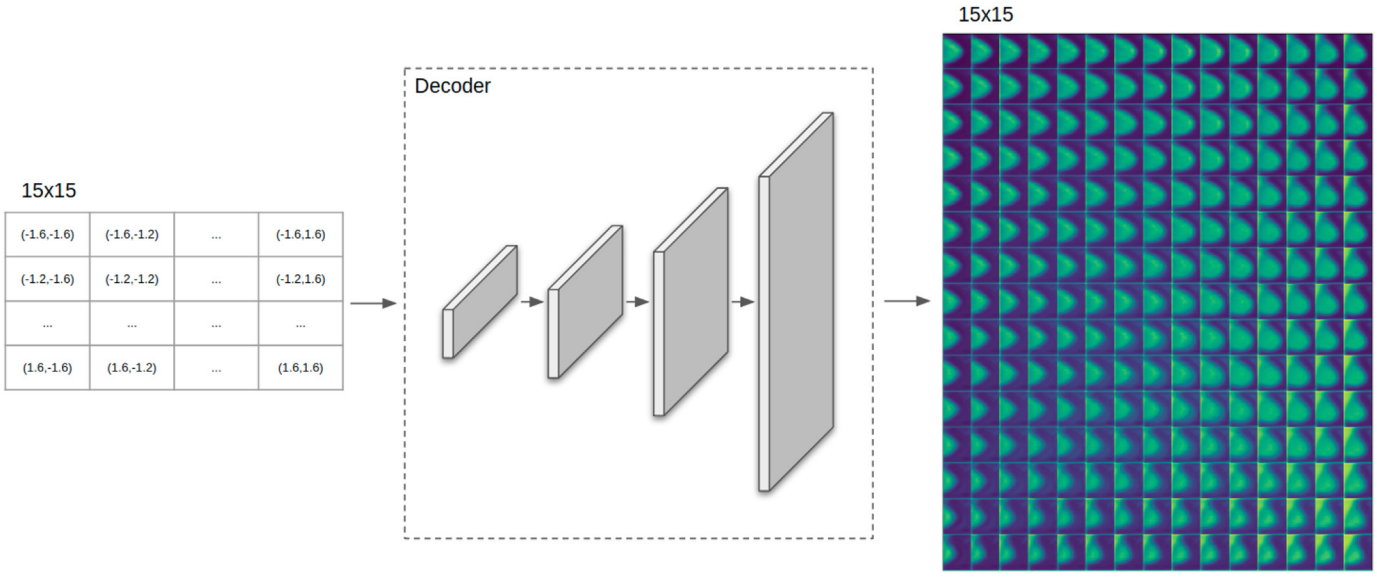

Fig. 7. Here the variational autoencoder decoder is decoupled and used as a stand-alone mammogram generator.

Table 1
Confusion matrix for the image acceptor.

Confusion Matrix	Mammogram	Brain MRI	Chest X-Ray	Gastrointestinal	COCO
Mammogram	408	0	0	2	0
Non-Mammogram	2	410	410	408	410

Although neural networks have a theoretical capacity to learn CLAHE-type processing, performing CLAHE during preprocessing reduces the processing burden of the neural network. The performance of CLAHE can be seen in Fig. 10 [25].

4.1.1. Digital database for Screening Mammography

The DDSM is a relatively large mammogram database, containing 10,412 mammograms [26]. 1,565 DDSM images depicting normal mammograms and 1,247 DDSM images depicting malignant mammograms, a total of 2,812 mammograms, were taken from the cancer volumes for training. Creating training data from the DDSM presented a couple of challenges. First, DDSM images are compressed in a format called lossless JPEG (LJPEG) and must be decompressed prior to being usable. Second, the DDSM does not directly provide bounding box coordinates for each abnormality. The coordinates of each abnormality bounding box must be approximated by a chain code which describes an outline around each abnormality [24].

For decompressing LJPEG images, the DDSM website provides the source code for a tool which converts LJPEG images to JPEG images. However, the source code was originally written for Unix (SunOS) and does not naturally compile on Linux (Ubuntu). Source code modifications were made as required such that the tool could compile and operate on Ubuntu Linux. Ultimately, the LJPEG conversion tool was not used directly, but a Python wrapper script by Wei Dong was leveraged to control the tool [24,27].

For addressing the second challenge, a short algorithm was written to convert the chain code to an approximate set of (x, y) coordinates for each abnormality bounding box. The chain code associated with each abnormality image provides starting (x, y) coordinates followed by a list of vectors which trace the outline of the abnormality. Each vector can only take on values of [0; −1], [1; −1], [1; 0], [1; 1], [0; 1], [−1; 1], [−1; 0], or [−1; −1] and provides a displacement relative to the end point of the previous vector [24]. As a first step to approximating the bounding box of each abnormality, all of the vectors were translated to be relative to the trace origin rather than the end point of the previous vector. That is, if the first and second trace vectors were [0; 1] and [1; 1] respectively, the first vector would remain as [0; 1] but the second vector would become [1; 2], the sum of the two vectors. Following this procedure for every vector in the chain code produced a new array of total displacement vectors. Examining each total displacement vector in this new array and extracting the minimum x-displacement value produced a value for the starting x-coordinate of the abnormality bounding box. The starting x-coordinate value was of course relative to the starting point of the chain code. Starting y-coordinate, ending x-coordinate, and ending y-coordinate values for each bounding box were found in a similar manner. These coordinates were scaled appropriately when the DDSM images were scaled down to a max dimension of 1,024 pixels in order to reduce the computational burden during training.

4.1.2. INbreast database

The INbreast database contains a total of 410 images of which only 124 images were used for validation. Images from the INbreast dataset have BI-RADS scores from 1 to 6 and only images with scores of 1, 5, and 6 were used. A BI-RADS score of 1 indicates a negative screening while scores of 5 and 6 indicate a high or certain probability that an identified abnormality is malignant. Images with scores of 2–4 were not used because they contain benign abnormalities or abnormalities of uncertain classification. For the INbreast images without a biopsy in place, our system could identify a malignant tumor but we have no way of knowing if the tumor is actually malignant or just benign. A biopsy is the only way to confidently identify breast cancer [23,28,29].

4.1.3. Dataset augmentation

Dataset augmentation effectively increased the number of training images by applying image transformations that did not change the classifications of the images. In the case of the mammogram dataset, each mammogram can be rotated 90° without changing the classification. Seven new training images can be generated from each original training image using the following transforms: 90° rotation, 180° rotation, 270° rotation, image mirror, image mirror followed by 90° rotation, image mirror followed by 180° rotation, and image mirror followed by 270° rotation. The dataset can be further augmented by shrinking or growing each image by 10%. Like rotation, minor image shrinking and growing will not change the image classification and will increase the number of training images to help prevent the neural network from simply memorizing the dataset. In other words, by increasing the number of training images, dataset augmentation can effectively reduce overfitting and increase overall validation accuracy for the model [30]. Random horizontal flips, vertical flips, rotations, scaling, and bounding box jittering all served as augmentation techniques for the DDSM images during training. After applying all possible rotation and flipping augmentation techniques, the training set is effectively 22,496 images. However, because the scaling and jitter operations can take on a range of possible values, it's difficult to quantify the total number of training images post-augmentation. Examples of

Table 2
Applications of deep learning in mammogram malignancy identification.

Author	Year	Data Source	Technique	Portion Analyzed	Result (AUROC)
This study	2018	INbreast	ResNet Faster R-CNN	Entire Image	0.951
Ribli et al. [8]	2017	INbreast/Dream	VGG Faster R-CNN	Entire Image	0.95/0.85
Akselrod-Ballin et al. [10]	2017	Private	VGG Faster R-CNN	Entire Image	0.72
Kooi et al. [34]	2017	Private	RF/CNN	Entire Image	0.941
Lotter et al. [35]	2017	DDSM	Multi-Scale CNN	Entire Image	0.92
Geras et al. [7]	2017	Private	Multi-View CNN	Entire Image	0.765
Shen [9]	2017	CBIS-DDSM/INbreast	3-Model Average	Entire Image	0.91/0.96
Dhungel et al. [36]	2017	INbreast	Multi-View ResNet	Entire Image	0.80
Jadoon et al. [5]	2017	IRMA	CNN-CT	Image Patch	0.855
Jiang et al. [37]	2017	BCDR-F03	Pre-trained GoogLeNet	Image Patch	0.88
Arevalo et al. [6]	2016	BCDR-F03	CNN	Image Patch	0.826

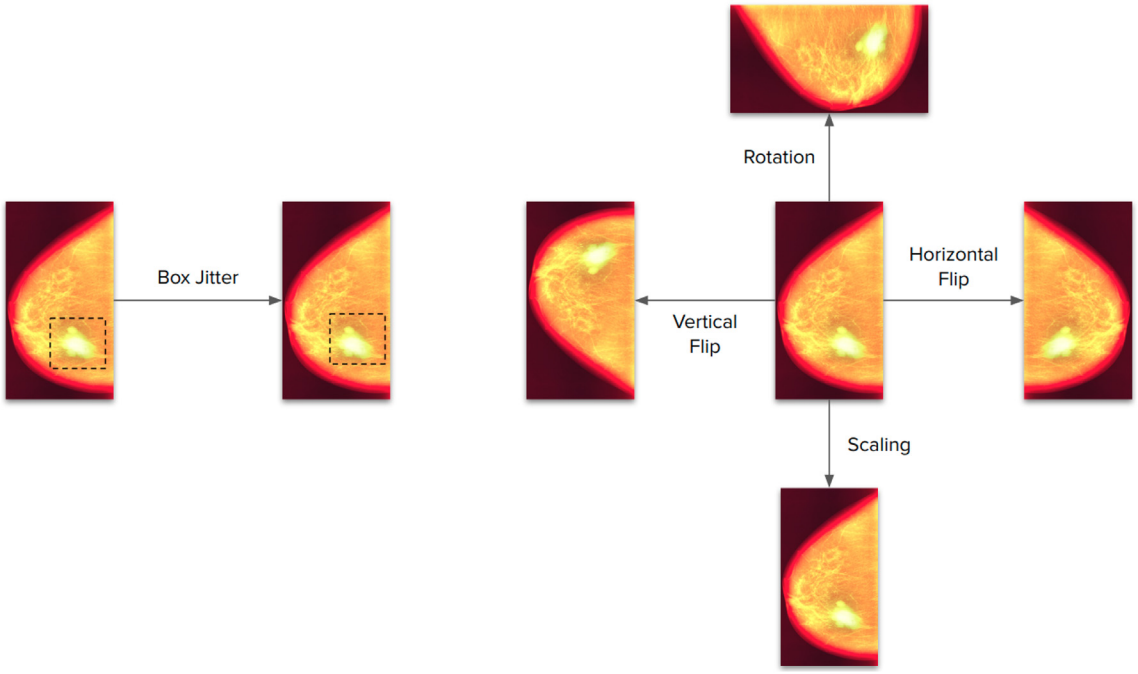

Fig. 8. Dataset augmentation techniques are shown in the figure above. These techniques were used to reduce overfitting during training of the tumor identifier.

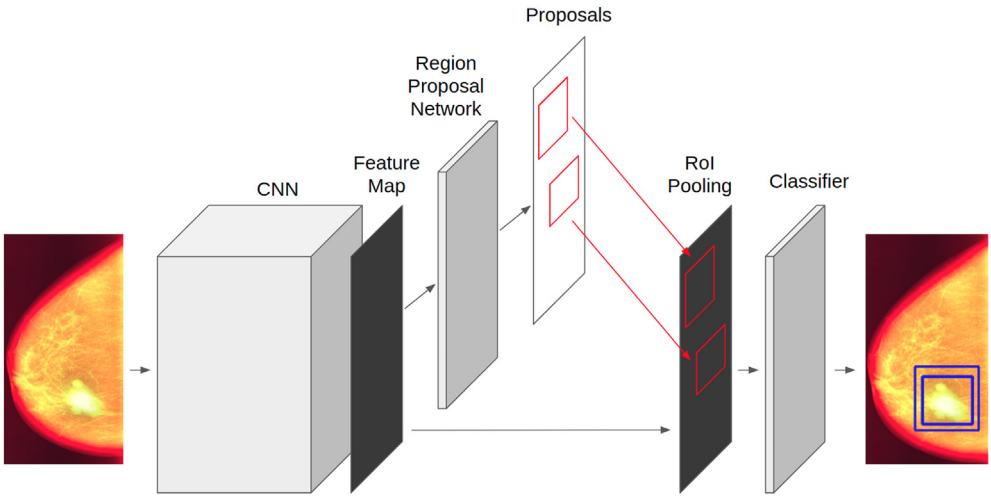

Fig. 9. Depicted is the architecture of the tumor identification network. This network architecture is popularly referred to as Faster R-CNN. The convolutional layers for feature extraction are based on ResNet-101.

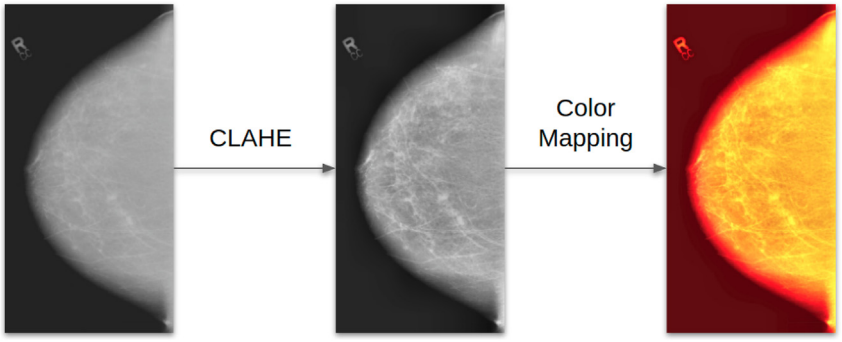

Fig. 10. Image preprocessing steps are shown in the image above. The image on the left is a mammogram before any preprocessing, the image in the middle is a mammogram after CLAHE, and the image on the right is a mammogram after CLAHE and color mapping. CLAHE and color mapping both enhance image details.

our dataset augmentation techniques can be found in Fig. 8.

4.1.4. ResNet-101 Faster R-CNN

The neural network architecture is commonly known as a ResNet-101 faster region-based convolutional neural network (Faster R-CNN). The term ResNet is an abbreviation for residual network. Residual networks are neural networks which are designed based on the highly successful residual learning framework published by Microsoft Research in 2016. The main idea behind residual nets is that it is much easier and more effective to learn adjustments from an identity function than to learn a completely new function from scratch. In a typical neural network, each set of layers attempts to learn some function $f(x)$. However, in a residual network, a set of layers known as a residual block is structured to learn the best $f(x) + x$. Therefore, residual blocks can easily learn functions similar to the identity function because it is presumably easy to learn $f(x) \rightarrow 0$ compared to $f(x) \rightarrow x$ [31]. Faster R-CNN, on the other hand, is an architecture proposed by Shaoqing Ren et al. [32] for quickly and accurately identifying objects of interest within a larger image. R-CNN architectures have a region proposal network (RPN) and a convolutional neural network (CNN) classifier. The RPN proposes regions of interest which the CNN then analyzes and labels. The innovation introduced by the Faster R-CNN architecture is that the RPN and final CNN layers can operate off of the same features which are extracted by the beginning layers of the CNN. In other words, Faster R-CNN combines both the RPN and CNN classifier into a single network instead of maintaining them as separate entities. Faster R-CNN has been shown to be more accurate and faster than previous R-CNN architectures [32]. Going back to the network used for this paper, the name ResNet-101 Faster R-CNN denotes that the network follows a Faster R-CNN architecture based on ResNet-101 layers. Although an Inception-ResNet-v2 Faster R-CNN architecture has been shown to produce higher accuracy than the ResNet-101 Faster R-CNN architecture, the ResNet-101 based architecture provides a very compelling balance between accuracy and required training time [11]. An overview of the architecture can be seen in Fig. 9. Google's TensorFlow Object Detection API was used for implementing the network for this paper [33].

4.2. Tumor identifier results

The neural network was trained using Google Cloud Platform computational resources and the performance of the neural network was evaluated on the INbreast dataset, achieving an area under curve (AUC) for receiver operating characteristic (ROC) curve of 0.951. A 95% confidence interval of 0.911–0.981 AUC was obtained by averaging the 95% confidence interval from 50 sets of 2,000 bootstrap samples. The ROC curve can be see in Fig. 11. Training required approximately 24 hours on an NVIDIA Tesla K80 while evaluation was performed on an Intel i7 processor and required 1 hour and 26 minutes. If multiple abnormalities were identified in an image, the highest estimated probability of malignancy across all abnormalities was used to represent the overall probability that the mammogram depicted a case of breast cancer. Fig. 12 depicts images that have been classified by the network proposed in this paper. The bounding box coordinates were output from the network and depict locations of abnormalities. A disadvantage of bounding boxes is that they visually depict the presence or absence of abnormalities in a binary fashion. That is, a bounding box is either present or absent — there is no visually inbetween state. Therefore, an alternative representation for visualizing the presence of abnormalities has been created in the form of a probability heatmap. The bottom 3 images in Fig. 13 contain probability heatmaps for abnormality localization. Each heatmap was created by collecting the 100 highest scoring bounding boxes, converting each bounding box to an ellipse of equivalent width and height, setting the pixel values of each ellipse equal to the score of the associated bounding box, and then summing together all 100 ellipses on a blank image with dimensions equal to the associated mammogram image. The generated heatmaps were then overlaid on associated greyscale mammogram images by setting the alpha channel of the heatmap to 32%. As can be seen in the lower right mammogram in Fig. 13, there is a slightly highlighted region which is not suspicious enough to generate a bounding box but could be of interest to a professional who is inspecting the mammogram for abnormalities.

4.3. Tumor identifier discussion

The AUC values presented by this paper are comparable to other AUC values listed in Table 2 of this report. It is difficult to directly compare results with some of the other papers because there is no standard dataset for training and evaluating mammogram classifiers. Also, several researchers have used datasets which are not publicly available. In addition to differences in dataset selection, the specific classification objective is not the same across all papers. While some researchers have tackled the problem of full mammogram image classification, other researchers have only attempted to classify smaller images of lesions as benign or malignant.

Although many studies are difficult to compare with, there are several worth discussing. One particularly notable system which was presented by Ribli et al. achieved an AUC of 0.95 on the publically available INbreast dataset. A caveat to this result, however, is that it was achieved with an ensemble of two networks. That is, two separate neural networks were independently trained and the score assigned to each image was the average of the scores output by the two networks. Although ensembling neural networks is an effective way to increase overall system accuracy, it can be computationally expensive and prevent direct comparison of results [8]. In comparison, this study used only a single network and likely could have performed even better if ensembling was leveraged. Shen [9] also used an ensemble of three neural networks to achieve an AUC of 0.91 on the CBIS-DDSM, but for comparison reports a 0.88 AUC for analysis by a single model. Unlike the original DDSM, the CBIS-DDSM used by Shen has 254 images removed from the dataset where a mass is not clearly visible. Shen achieved a 0.96 AUC on the INbreast dataset and included all images except those with BI-RADS readings of 3. However, Shen used a portion of the INbreast dataset for training prior to validation on the remaining images [9]. For the network proposed in this paper, no training was performed on the INbreast dataset prior to validation. Also, this study used the original DDSM rather than the CBIS-DDSM because of accessibility. Lastly, Lotter et al. [35] only evaluated performance on the DDSM, but achieved a substantial AUC of 0.92. Lotter et al. performed considerable dataset augmentation, producing up to 900K image patches for one particular stage of training [35]. Although we did not perform this level of augmentation, dataset augmentation among other strategies could improve the results described in this paper.

In future work, accuracy of the malignancy identification neural network or image acceptance neural network could be improved via hyperparameter optimization. Hyperparameters refer to different aspects of the neural network such as learning rate, network structure, gradient descent technique, or dropout rate which are established prior to actual training. Almost no optimization was attempted in any of these four hyperparameters. Therefore, optimizing these and other hyperparameters could offer substantial improvements to the model's accuracy [38].

As mentioned previously, another strategy for optimization is dataset augmentation. Although the dataset was augmented via random horizontal flips, vertical flips, rotation, scaling, and bounding box jitter, additional techniques such as artificial image generation, random contrast or saturation changes could also be used to augment the training data [26]. Similar to dataset augmentation, validation augmentation can also increase validation accuracy. In one form of validation augmentation, an image and the mirror of the image are both evaluated by the model. The average of the two evaluation scores is

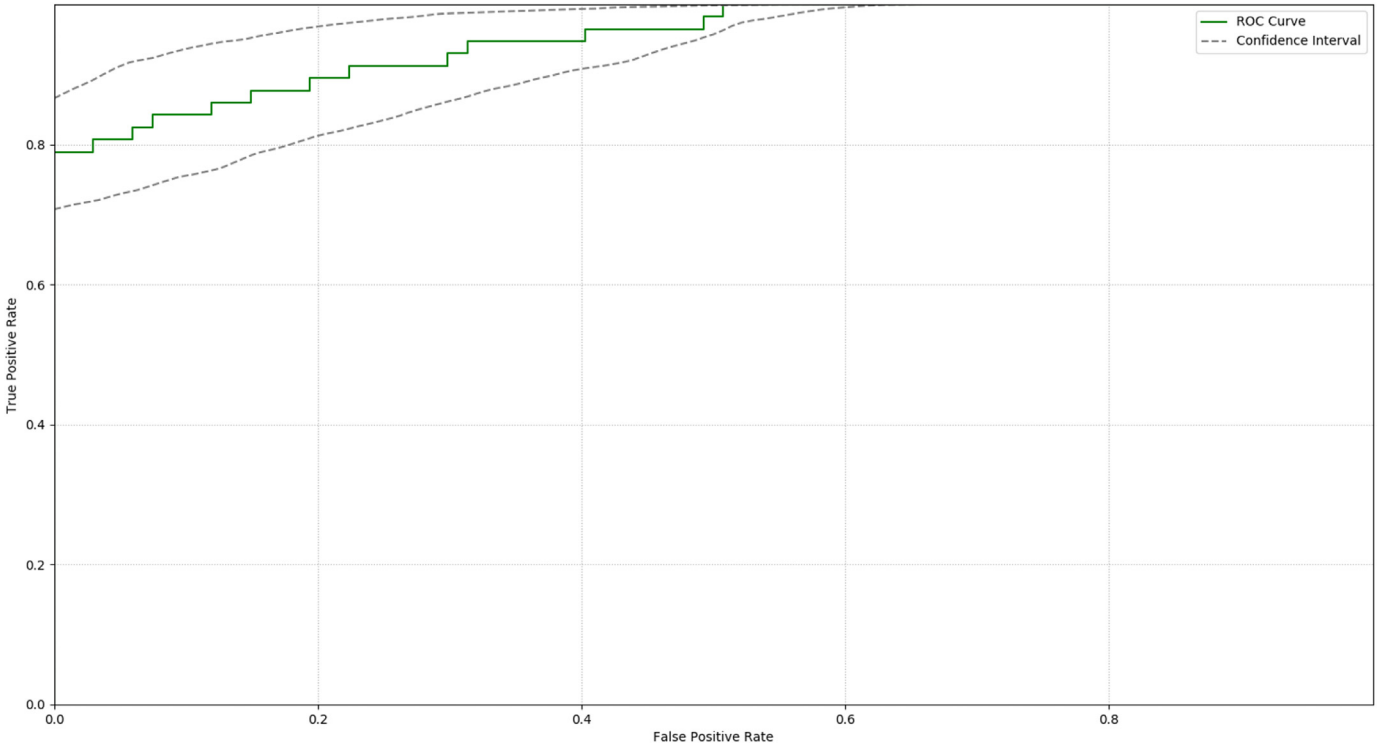

Fig. 11. The solid green line shows the ROC curve (sensitivity versus 1 - specificity) obtained for the INbreast dataset (0.951 AUC) and the dashed grey lines show the approximated 95% confidence interval.

then reported as the overall evaluation score. However, the image to be evaluated may be augmented by any of the augmentation methods described for dataset augmentation [8].

Aside from dataset augmentation, adjustments in scaling could increase accuracy in another way. The images used for training were DDSM images that had been downscaled to a max dimension of 1,024 pixels. Although this downscaling was intended to reduce the computational burden of training, the downscaling may have removed significant details and ultimately harmed the development of the classifier. In future research, full size images with only augmentation related scaling could be used to train a more robust network.

In addition to removing downscaling from the training images, the training dataset could be improved by removing questionable images. There is a publically available subset of the DDSM known as the Curated Breast Imaging Subset of DDSM (CBIS-DDSM). According to the dataset's summary, the original DDSM contains 254 images where a mass is not clearly visible. These 254 questionable images have not been included in the CBIS-DDSM. Removing these images from the DDSM dataset or performing training with the CBIS-DDSM could improve both training and validation accuracy [39].

Lastly, validation accuracy could be improved in future models via highly relevant transfer learning. Transfer learning is where a model or parts of a model trained for a specific task are reused for a new task. Transfer learning is very popular in image recognition applications because many low-level recognition objectives (i.e. edge, corner, or line recognition) are universal across many if not all image recognition tasks. Therefore, the layers of a deep convolutional model that have been trained on millions of images to recognize basic image features could be reused for the task of mammogram classification. Although the neural network used in this paper was pre-trained using the COCO

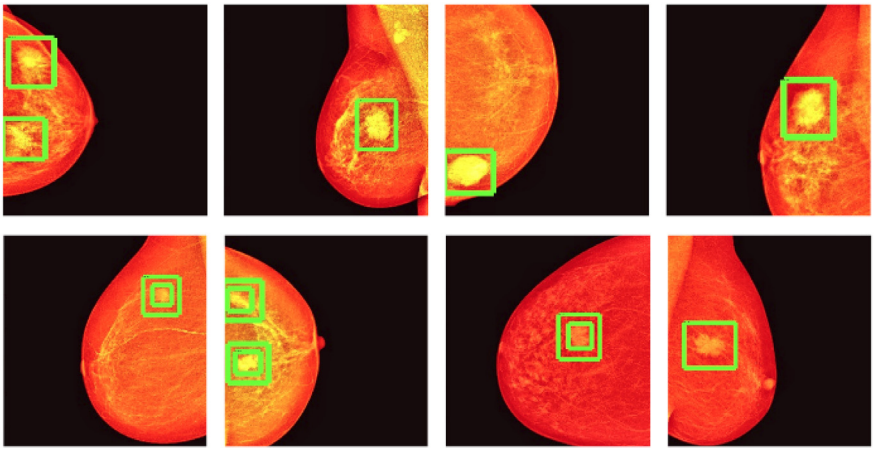

Fig. 12. Examples of tumors identified by the algorithm are shown in the images above. Green bounding boxes identify the location of tumors where stacked boxes represent uncertainty in the tumor size.

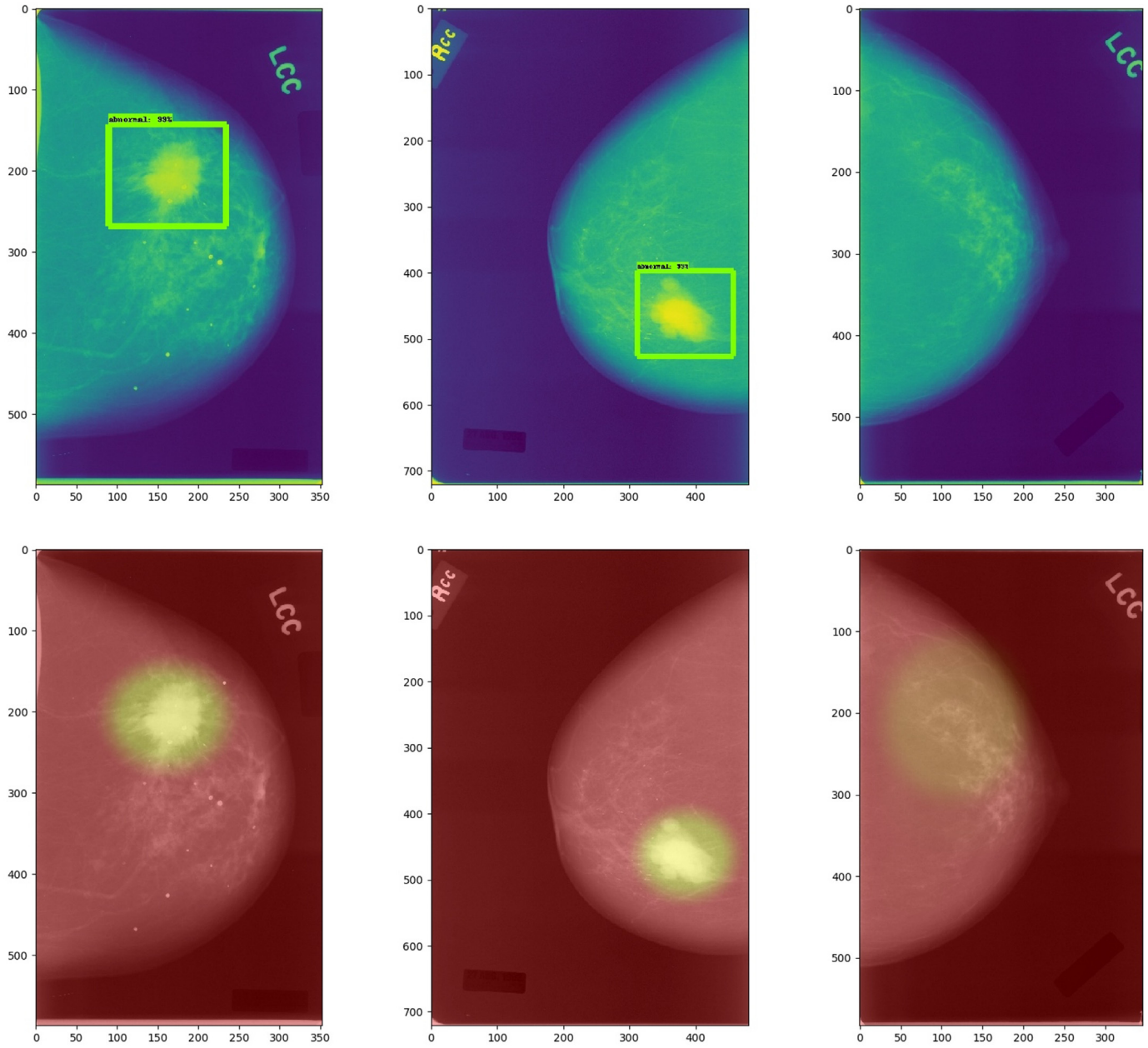

Fig. 13. In the top left and middle images, abnormalities have been identified as indicated by the green bounding boxes. In the top right image, no abnormality has been identified. In the bottom images, a heatmap is used to depict the probability that an abnormality is present.

dataset, pre-training with a dataset of assorted medical x-rays would perhaps be more effective and meaningful. Although low level image features such as edges are universally relevant across image recognition tasks, higher level image features may be less universally relevant. Simply training with a large database of labeled mammograms would be the most ideal. With all of these strategies in mind, the existing system could be made even better [38].

5. Conclusion

We believe we have presented a novel and comprehensive solution to early breast cancer detection which is both fast and accessible. The telemedical service provides both doctors and patients easy access to state-of-the-art malignancy identification software. In addition, the image acceptor addresses invalid user input, an issue that is common across all openly accessible web services. We believe the work presented here will be impactful not only for mammogram analysis but also for all kinds of medical image processing.

The image acceptor achieved very respectable results, misclassifying only 2 out of 1,640 non-mammograms and 2 out of 410 mammograms from the validation dataset. Since this system has only been trained on mammograms, we have high confidence that it will be robust to any class of non-mammogram image.

The tumor identifier also performed well, similar to other state-of-the-art tumor classifiers. Although validation of this classifier was performed on a relatively small dataset, bootstrapping 100,000 samples provided a 95% confidence interval of 0.911–0.981 AUROC. In addition, as alike deep network architectures have been widely leveraged in general image recognition and even breast tumor identification, we are confident that our classifier provides a robust solution.

Conflict of interest

None.

Acknowledgments

The authors would like to thank Dr. Christopher Simmons from the University of Texas at Dallas for his assistance in hosting the web application on a publicly available server. The DDSM used for training was provided courtesy of the University of South Florida. The INbreast database used for validation was provided courtesy of the Breast Research Group, INESC Porto, Portugal. The authors also acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources which trained early malignancy identification prototypes. URL: http://www.tacc.utexas.edu.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (6) (2018) 394–424.

[2] P. Salzmann, K. Kerlikowske, K. Phillips, Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age, Ann. Intern. Med. 127 (11) (1997) 955–965.

[3] N. Perry, M. Broeders, C. de Wolf, S. Törnberg, R. Holland, L. von Karsa, European guidelines for quality assurance in breast cancer screening and diagnosis. Summary document, Ann. Oncol. 19 (4) (2008) 614–622.

[4] C.A. Beam, P.M. Layde, D.C. Sullivan, Variability in the interpretation of screening mammograms by us radiologists: findings from a national sample, Arch. Intern. Med. 156 (2) (1996) 209–213.

[5] M.M. Jadoon, Q. Zhang, I.U. Haq, S. Butt, A. Jadoon, Three-class mammogram classification based on descriptive cnn features, Biomed. Res. Int. 2017 (2017).

[6] J. Arevalo, F.A. González, R. Ramos-Pollán, J.L. Oliveira, M.A.G. Lopez, Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. 127 (2016) 248–257.

[7] K.J. Geras, S. Wolfson, S. Kim, L. Moy, K. Cho, High-resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks, (2017) arXiv preprint arXiv:1703.07047.

[8] D. Ribli, A. Horváth, Z. Unger, P. Pollner, I. Csabai, Detecting and classifying lesions in mammograms with deep learning, Sci. Rep. 8 (1) (2018) 4165.

[9] L. Shen, End-to-end Training for Whole Image Breast Cancer Diagnosis Using an All Convolutional Design, (2017) arXiv preprint arXiv:1708.09427.

[10] A. Akselrod-Ballin, L. Karlinsky, S. Alpert, S. Hashoul, R. Ben-Ari, E. Barkan, "A Cnn Based Method for Automatic Mass Detection and Classification in Mammograms," *Computer Methods In Biomechanics And Biomedical Engineering: Imaging & Visualization*, (2017), pp. 1–8.

[11] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al., Speed/accuracy Trade-Offs for Modern Convolutional Object Detectors, (2016) arXiv preprint arXiv:1611.10012.

[12] Mpl colormaps, https://bids.github.io/colormap/ , Accessed date: 11 November 2017.

[13] R. Chalapathy, E. Toth, S. Chawla, Group Anomaly Detection Using Deep Generative Models, (2018) arXiv preprint arXiv:1804.04876.

[14] P. Perera, V.M. Patel, Learning Deep Features for One-Class Classification, (2018) arXiv preprint arXiv:1801.05365.

[15] X. Hou, L. Shen, K. Sun, G. Qiu, Deep feature consistent variational autoencoder, Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, IEEE, 2017, pp. 1133–1141.

[16] C. Doersch, Tutorial on Variational Autoencoders, (2016) arXiv preprint arXiv:1606.05908.

[17] A.B.L. Larsen, S.K. Sønderby, H. Larochelle, O. Winther, Autoencoding beyond Pixels Using a Learned Similarity Metric, (2015) arXiv preprint arXiv:1512.09300.

[18] K. Pogorelov, K.R. Randel, C. Griwodz, S.L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P.T. Schmidt, M. Riegler, P. Halvorsen, Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection, Proceedings of the 8th ACM on Multimedia Systems Conference, Ser. MMSys'17. New York, NY, USA, ACM, 2017, pp. 164–169 [Online]. Available: http://doi.acm.org/10.1145/3083187.3083212.

[19] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 3462–3471.

[20] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, Q. Feng, Enhanced performance of brain tumor classification via tumor region augmentation and partition, PLoS One 10 (10) (2015) e0140381.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[22] K. Ganesan, U.R. Acharya, C.K. Chua, C.M. Lim, K.T. Abraham, One-class classification of mammograms using trace transform functionals, IEEE Trans. Instr. Measur. 63 (2) (2014) 304–311.

[23] Inbreast database, http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database , Accessed date: 4 November 2017.

[24] Digital database for screening mammography, http://marathon.csee.usf.edu/Mammography/Database.html , Accessed date: 5 August 2017.

[25] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J.B. Zimmerman, K. Zuiderveld, Adaptive histogram equalization and its variations, Comput. Vis. Graph Image Process 39 (3) (1987) 355–368.

[26] M. Kim, J. Zuallaert, W. De Neve, "Towards Novel Methods for Effective Transfer Learning and Unsupervised Deep Learning for Medical Image Analysis," in Doctoral Consortium (DCBIOSTEC 2017), (2017), pp. 32–39.

[27] W. Dong, Ljpeg decompressor script, https://github.com/aaalgo/ljpeg , Accessed date: 5 August 2017.

[28] M.M. Eberl, C.H. Fox, S.B. Edge, C.A. Carter, M.C. Mahoney, Bi-rads classification for management of abnormal mammograms, J. Am. Board Fam. Med. 19 (2) (2006) 161–164.

[29] F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, A dataset for breast cancer histopathological image classification, IEEE Trans. Biomed. Eng. 63 (7) (2016) 1455–1462.

[30] M.A. Nielsen, Neural Networks and Deep Learning, Determination Press, 2015, http://neuralnetworksanddeeplearning.com/.

[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[32] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[33] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, (2015) software available from: tensorflow.org. [Online]. Available: http://tensorflow.org/.

[34] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C.I. Sánchez, R. Mann, A. den Heeten, N. Karssemeijer, Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312.

[35] W. Lotter, G. Sorensen, D. Cox, A Multi-Scale Cnn and Curriculum Learning Strategy for Mammogram Classification, (2017) arXiv preprint arXiv:1707.06978.

[36] N. Dhungel, G. Carneiro, A.P. Bradley, Fully automated classification of mammograms using deep residual neural networks, Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, IEEE, 2017, pp. 310–314.

[37] F. Jiang, H. Liu, S. Yu, Y. Xie, Breast mass lesion classification in mammograms by transfer learning, Proceedings of the 5th International Conference on Bioinformatics and Computational Biology, ACM, 2017, pp. 59–62.

[38] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.

[39] Curated breast imaging subset of the digital database for screening mammography, https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM#1b292186190048e88292b80331d47e8c , Accessed date: 7 November 2017.