International Journal of Engineering in Computer Science

E-ISSN: 2663-3590
P-ISSN: 2663-3582
IJECS 2022; 4(1): 49-56
Received: 15-03-2021
Accepted: 20-04-2021

Viswanatha Reddy Allugunti
Chief Solutions Architect
Arohak Inc, USA

Breast cancer detection based on thermographic images using machine learning and deep learning algorithms

Viswanatha Reddy Allugunti

DOI: https://doi.org/10.33545/26633582.2022.v4.i1a.68

Abstract
According to the latest data, breast carcinoma is the most prevalent kind of cancer in the world, and it is responsible for the deaths of almost 900 thousand people each year. If the disease is detected at the early stage and diagnosed properly, it can improve the chance of positive outcomes, thus reducing the fatality rate. An early diagnosis in fact can help in preventing it to spread and saves the premature victims from obtaining it. When trying to distinguish among benign and malignant tumors, as well as when trying to draw conclusions about mild and advanced breast cancer, researchers who study cancer encounter a number of challenges. The identification of all tumors is accomplished through the application of machine learning, which makes use of algorithms that are able to locate and recognize patterns. All of them, however, revolve around the concept of "binary grouping," as was mentioned earlier (malignant and benign; no-cancer and cancer). In this study, we propose a Computer-aided Diagnosis (CAD) method for the identification and diagnosis of patients into 3 classes (cancer, no cancer, and non-cancerous) under the management of a database. CAD is an abbreviation for computer-aided diagnosis. The Convolutional Neural Network (CNN), the Support Vector Machine (SVM), and Random Forest are all remarkable classifiers (RF). Convolution Networks, Support Vector Machines (SVM), and Random Forest are the three effective classifiers that we look into and analyses for the classification stage (RF). In addition to this, we investigate the impact of the mammography pictures being pre-processed in advance, which allows for a higher success rate in categorization.

Keywords: Deep learning, breast cancer; thermal imaging; machine learning

1. Introduction

Breast cancer is the most frequent kind of cancer in women, as stated by the Centers for Disease Control and Prevention (CDC), a reliable source. The likelihood of surviving breast cancer varies greatly depending on a variety of factors. The form of tumor a woman has and the stage of the disease at the time she is diagnosed with it are two of the most crucial aspects. Cancer that originates in the breast cells is known as breast cancer. In most cases, the cancer begins to develop in one of the two areas of the breast known as the lobules or the ducts. Within your breast, cancer can also develop in the fatty tissue, known as adipose tissue, as well as the fibrous connective tissue. Cancer cells that are not under control may frequently infect other healthy breast tissue, and they may also travel to the lymph nodes located beneath the arms.

Breast cancer, according to medical professionals, is caused by the uncontrolled growth of abnormal cells in the breast, and these cells are said to have spread in growth like Meta Size from the breast to the lymph nodes or other areas of the body as well. In order to avoid the repercussions of the subsequent phase, it is vital to find these undesired cells as soon as possible and put a halt to their proliferation as soon as feasible. If a tumour is found, the first thing a doctor will do is determine whether or not the growth is benign by determining whether or not it can divide into two categories: benign and malignant. Because the strategies for treating and preventing both types of cancers are distinct from one another. Malignant cells are the ones that can become malignant and spread to other regions of the body, whereas benign cells do not develop into cancer and do not spread. The difficulty with all of this illness is because there is no screening device of that kind and quality that really can detect cancer in its early stages. If there were a device like this one, a patient would indeed be ready to initiate medication as quickly as possible and work toward preventing the growth of undesirable cells or malignancies. However, there is currently no such machine.

Correspondence
Viswanatha Reddy Allugunti
Chief Solutions Architect
Arohak Inc, USA

If you have any kind of ailment, getting an early diagnosis is almost always the key to successfully treating it. The majority of people are unable to diagnose their illness before it progresses to a chronic stage. It contributes to a rise in the number of people who are passing away all over the world. When detected in its earlier stages, breast cancer is one of the diseases that has a chance of being cured. This is because earlier detection of the disease prevents it from spreading to other parts of the body.

The absence of prognostic models makes it difficult for medical professionals to devise treatment strategies that have the potential to lengthen a patient's overall survival time. Therefore, time is required to discover the strategy that produces the least amount of error in order to improve accuracy. Because the currently available methods to identify breast cancer, such as mammograms, ultrasounds, and biopsies, take a significant amount of time, there was a demand for a computerised diagnostic system that utilised the technique of machine learning. This methodology makes use of algorithms that speed up the process of classifying the tumour, improve the accuracy with which cells are located, and shorten the amount of time required to do so.

In recent years, thermography has become an increasingly widespread method, particularly for the detection of cervical cancer [12]. This is because of the appealing realities from its own relatively safe invention, in addition to the chance of future upgrades made possible by cutting-edge technical improvement. The current research being conducted in this area is to arrive at a tumour outcome that is more definitive and can be agreed upon by a large number of people, and which can be utilised as a recommendation for breast cancer screening. In a similar vein, overcoming the recently formed barriers that are posed by the time-consuming screening procedure, particularly when photo preparation is necessary.

The field of thermographic imaging as a science and the applications it can serve have been given new life because to ongoing innovations in the field. Probably one of the best applications of thermography is the screening for breast cancer, which is performed using the technique. In spite of this, thermography has not yet been validated as the method of choice for carrying out this specific task. In addition, given that thermography is not exactly a risk-free operation, most medical professionals would rather have the results of a mammogram rather than those of a thermograph. As a consequence of this, thermal imaging screening mammography has the potential to develop into a viable alternative choice if it is enhanced to a significant degree. The primary topic that needs to be covered in this conversation is how to imagine getting ready to execute the task. The findings of this study suggest that a Convolutional Neural Network, or CNN, should be utilised for thermal imaging testing and treatment in order to mitigate the drawbacks described before.

2. Related work

Imaging and genomics-based methods of breast cancer detection have been the subject of a significant number of investigations. Moreover, to the best of the knowledge, there has not been any research carried out that incorporates both of these methods.

The authors of [1] provided a comprehensive summary of the numerous methods that are used to diagnose breast cancer via histological image analysis (HIA). These methods are based on various designs of convolutional neural networks (CNN). The authors organised their work into categories according on the dataset that was used. They put it in reverse chronological sequence, starting from the most recent event. According to the findings of this research, ANNs were utilised for the first time in the field of HIA around the year 2012. ANNs and PNNs were the algorithms that were utilised the most frequently. On the other hand, the majority of the work in feature extraction made use of textural and morphological traits. It was abundantly obvious that Deep Convolutional Neural Networks were highly beneficial for the early identification and treatment of breast cancer, which ultimately led to more successful therapy. Numerous algorithms were utilised in the process of making predictions for non-communicable diseases (NCDs).

The authors of [2] examined and compared the efficacy of a number of different classification methods. Using eight different classification methods and a 10-fold cross validation approach, the classification techniques were run on eight different NCD datasets. As a measure of accuracy, area under the curve was applied to these and evaluated. According to the authors, the NCD datasets contain noisy data as well as qualities that are not relevant. KNN, SVM, and NN all showed resilience in the face of this noise. In addition, they indicated that the issue of irrelevant attribute can be resolved by employing some pre-processing procedures, which would result in an increase in the rate of accuracy.

Natural inspiration computing (NIC) methodologies have been proposed and implemented in the process of diagnosing a variety of human conditions. Five insect-based NIC diagnostic algorithms were presented by the authors of [3], who discussed their utility in detecting diabetes and cancer. According to the authors, it identified several tumours well (breast, lung, prostate and ovarian). Combining directed ABC with neural networks helped diagnose breast cancer. In addition to this, the writers came up with a method that is extremely efficient for diagnosing diabetes and leukaemia. They came to the conclusion that the combination of NICs with the other classification methods generates findings that are both more accurate and promising. They underlined the need for additional work to be done in order to detect the various stages of diseases and diabetes.

The authors of [4] presented evidence that indicated the usefulness of NNs in the categorization of cancer diagnoses, particularly in the earlier stages of the disease. Their research demonstrates that the number of NNs have demonstrated some level of potential in the detection of malignant cells. However, in order to pre-process the images, the imaging method demands a significant amount of processing resources.

The authors of review article [5] discussed a variety of machine learning, deep learning, and data mining methods that are connected to breast cancer prediction. There were a total of 27 publications in machine learning, 4 articles in addressing the challenges related, and 8 articles in convolutional neural networks that were examined in this review of breast cancer research studies. The authors observed that the majority of the publications utilised imaging, but only a few papers utilised genetics in their research. The support vector machine (SVM), the decision tree, and the random forest were the primary algorithms utilised in the genetic analysis of breast cancer. Imaging approaches, on the other hand, made use of a variety of algorithms, such as CNNs and Naive Bayes.

But at the other extreme, the researchers in [6] concentrated their attention on the mutation of genes as a method for diagnosing breast cancer. They mentioned that now the reverse genetics classification phase intends to do gene annotation, gene discovery, and gene mutation detection in order to determine whether or not a malignancy is present. They came to the conclusion that numerous approaches, such as regression, probabilistic models, SVMs, neural networks, and deep learning, may be applied. They also stated the various options that are accessible to capture the connection among nucleotides and feature extraction. This is due to the fact that DNA sequencing involves a big amount of information in the shape of a strings sequence.

The authors of [7] reviewed current works that applied deep learning to breast cancer using a variety of imaging modalities. They structured these investigations by focusing on the datasets, architecture, applications, and evaluations involved. They concentrated on developing deep learning frameworks for breast imaging using three different modality (ultrasound, mammography and MRI). They wanted to deliver findings that were up to date with regard to breast cancer imaging by applying DLR-based CAD systems, and that was the focus of their effort. Their research incorporated the use of confidential datasets and CNNs for categorization. Following an analysis of these surveys, my contribution will consist of doing research into genetic sequencing and imaging simultaneously in order to forecast breast cancer and obtain more information that can assist in the early detection and treatment of breast cancer. In addition to this, we will offer suggestions to researchers who are interested in carrying out investigations in this field.

Using a variety of machine learning methods, the authors of [8] suggest an instinctual method for classifying mammography pictures as benign, malignant, or normal. An investigation into the similarities and differences among Support Vector Machines, Convolutional Neural Networks, and Random Forests is carried out. The findings of the simulation led researchers to the conclusion that CNN is the most effective classifier since it leads to the intuitive categorization of digital mammograms by employing morphological and filtering operations.

In [9] the author uses Dr. William H. Walberg's dataset from UW Hospital. This collection was used to practise data visualisation and machine learning techniques like logistic regression, k-nearest neighbours, SVM, naive Bayes, decision tree, random forest, and rotation forest. The programming languages R, Minitab, and Python were selected for application to these different machine learning strategies and visualisations. An investigation into the similarities and differences between each of the procedures was carried out. The results that were obtained using the logistic regression model that contained all of the features exhibited the maximum level of classification accuracy (98.1%), as well as the proposed method revealed an improvement in accuracy performances.

In [10], we carried out an investigation into the similarities and differences between SVM, Logistic Regression, Naive Bayes, and Random Forest. The comparison is carried out with the use of the dataset on breast cancer in Wisconsin. According to the findings of the tests that were carried out, the Random Forest algorithm had the best accuracy (99.76 percent) while also exhibiting the lowest error rate. The Anaconda Data Science Platforms was utilised in order to carry out all of the tests in a setting that was simulated.

The authors [11] suggested a method for breast cancer that differentiates between the many subtypes of breast cancer. This method relies on the Wisconsin Diagnosis and analysis and Prognostic Breast Cancer datasets for feature selection. It then uses a neural network approach to classify the various forms of breast cancer, paying particular attention to the MLP and the back - propagation neural RBF. The input layer of the neural network is represented by the nine features that are included in this data set. The features that are input will be categorised by the neural network into two distinct types of cancer (benign and malignant). The approach that was devised and tested on the database ended up resulting in a 97 percent recurrence rate of classification when RBF neural network was used.

In order to construct an ensemble model for the prediction of the severity of breast masses, the authors [12] tested and compared two distinct Bayesian classifiers, namely tree-augmented Naive Bayes and Markov blanket estimating networks. The purpose of the suggested algorithm was to assist medical professionals in making decisions regarding whether or not a breast biopsy should be performed on a worrisome lesion based on the results of a mammogram. The authors have discovered that the Based on bayesian classifiers are a viable alternative to several other methods that can be used in medical applications.

Authors [13] decide to follow Bayesian networks (BN) in the field of emergency medicine, where BN have been managed to find to be an effective methodology due to their powerful symbol, handling of uncertainty, and where various possibilities are possible based on the evidence that is given. Bayesian networks are found to be an effective methodology because of their own symbolic representation.

3. Proposed system

In this research, we describe a method that is both accurate and efficient for fragmenting thermal imaging breast images and diagnosing breast cancer so that the images can be categorised as normal or pathological, which translates to something like that without disease or with cancer. In order to do analysis and classification on the segmented thermographic images, we suggested utilising a convolutional neural network (CNN).

For this approach, image input variables are divided into pre-expanded RGB and Gray channels, which are merged with an autonomous image denoising and classification; the final outcome is going to feed both processes to the breast image analysis and feature extraction network, a two-extraction including one nest, which determines if the image is benign or malignant.

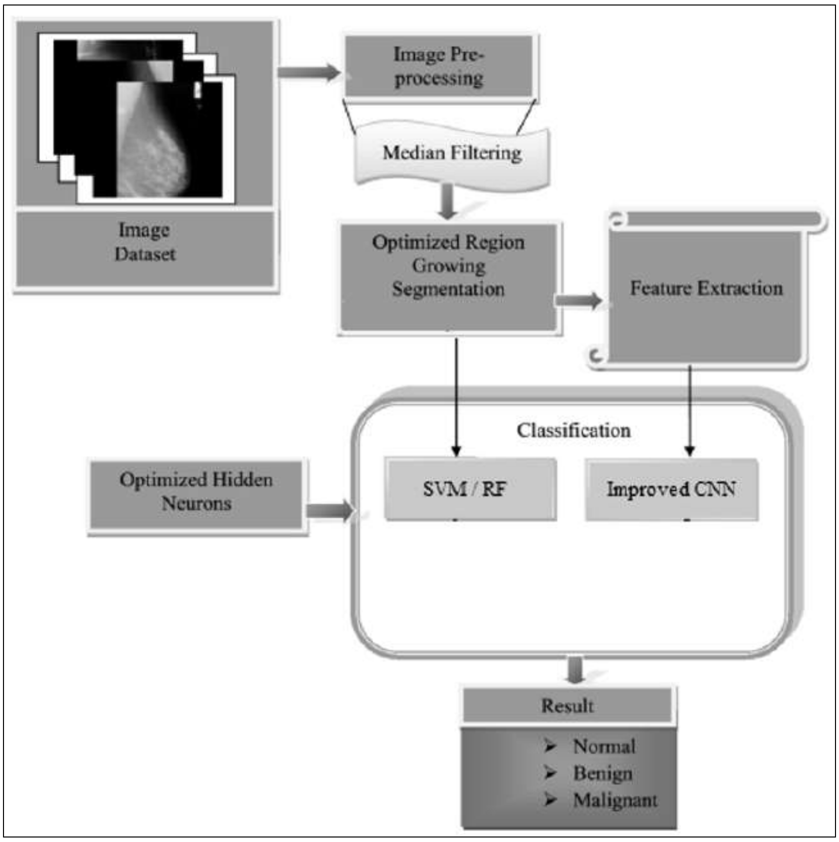

Fig 1: System Architecture

The measures that must be taken in order to put our proposed system into action are as follows.

1. Image dataset should be input into the system.
2. Image quality can be improved by performing pre-processing on the image.
3. From the input picture dataset, which serves as the basis for the generation of the training file, several features are retrieved.
4. The enhanced CNN classification method is then used to both the newly created training file dataset and the new test input images.
5. In addition, we employ SVM and Random Forest for the purpose of comparative analysis.
6. The cancer detection that results from using the CNN algorithm depends on whether the input test reveals normal, benign, or malignant cells.
7. At the very end, a graphical evaluation is carried out in order to assess how well the suggested system functions.

3.1 Implementation details

The approaches that are described in this paper are intended to evaluate infrared images of persons who are healthy as well as those who have cancer; following this, there are many attributes to extract distinct classes of the patients. The remaining photographs are organised into specialised classification folders, from which only fragments are drawn for the purpose of teaching classifiers while the remaining folders and files are analysed for the purpose of evaluating classifier performance. It is now possible to use it to recognise any new photo, despite the fact that it was taught with certain content classifications.

A. Dataset

The dataset that was utilised included photos of approximately over 150 patients, either with or without breast cancer, totalling over 1000 photographs. These images were discovered on the website Kaggle. Only the frontal pictures with the arms lifted were utilised for this particular piece of work because the other poses yielded inconsistent results. After that, the area of interest (ROI), also known as the zone that solely contained the patient's breast, was isolated from each image in the database using a software programme. Additionally, the files were scaled to have dimensions of 128x128 pixels each.

B. Image Pre-Processing

The pre-processing stage of picture production is essential to the production of clear and unmistakable images. The stage of picture pre-processing makes it possible to proceed to the step of categorising. The data augmentation process was utilised as the first step.

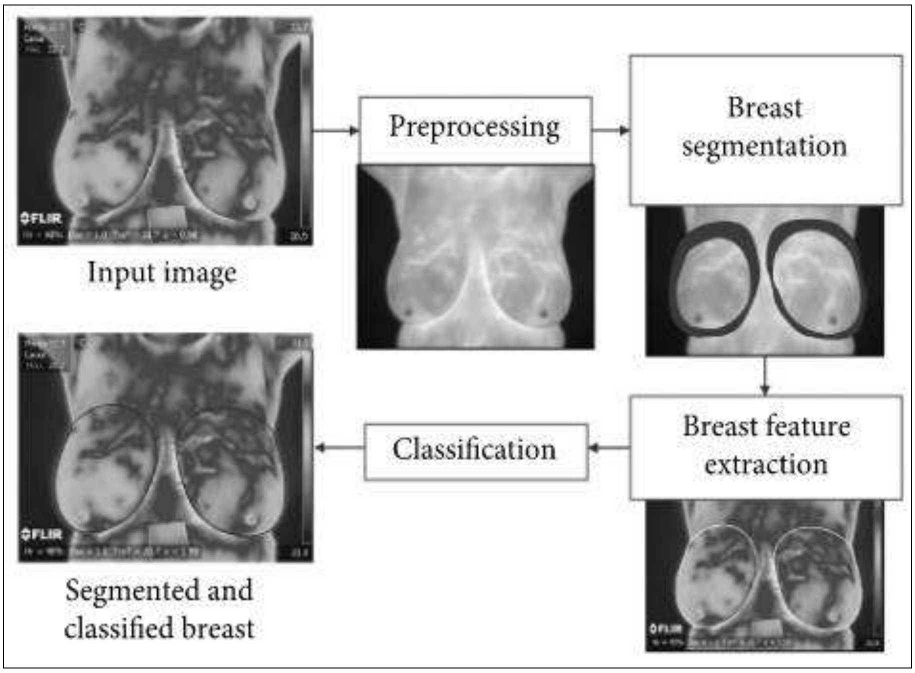

Fig 2: Block Diagram

This stage is responsible for contributing to the overall increase in the size of the dataset because it involves applying several conversions to the initial input. The input was iterated using a number of different conversions, including rotations, symmetries, and translations. The following is a description of the processes done for the pre-processing of the enhancement:

1) **Conversion:** The picture can be reduced to a given number of pixels aligned in a particular direction.
2) **In order to properly centre the images:** The columns and rows that were initially on the margins of each image had to be removed.

This was done by cutting off the excess columns and rows. As a consequence of this, photos can be obtained in a variety of sizes. After that, the rows and columns are trimmed down to the exact number, and the total number of images is tallied. After that, the files are levelled so that they are all the same size before being enlarged. After the pre-processing step, the randomised photos for healthy patients and sick patients are selected according to their degree of transparency.

C. Segmentation

Before beginning the process of extracting and classifying high-specific high-resolution data, it was necessary to finish the goal and segmentation operations first. This enabled the best possible outcomes. Because spectral problems were more prevalent, delineation suffered, resolution in the segmented files with a higher resolution was reduced, and image information was lost. To improve upon this, the object-oriented picture segmentation technique was applied. This technique removed the salt and noise from the image while simultaneously boosting the image's precision through the application of spectral signatures and the shapes of the objects themselves.

D. Feature Extraction

The extraction of attributes is helpful in scalability of the procedure as well as the production of more meaningful datasets consisting of larger and better quality images. There are alternative methods of attribute extraction that require the images to be processed first, however the CNN can directly obtain the properties of the input data without any additional processing. This approach of extracting characteristics from an image uses convolution as the primary tool, which enables the segmentation of a picture into its component parts. This is the case due to the fact that, in most instances, extracted features do not move, which means that the measurements and characteristics of one half of the image are equal with those of the other side of the image. This is the case because natural images do not move.

E. Classification

During classification process, a destabilization matrix were utilised to provide a wide description as to how the classifier carried out the classification operation. This was done so that the results of the classification could be more accurately interpreted. This explanation was used as part of the classification process. The performance of certain individuals is typically the primary focus of the overview. The matrix has been loaded with a test set of pre-defined tags that have already been determined to be correct after being considered. The data was then processed by a CNN classifier, resulting resulted in the production of predictions.

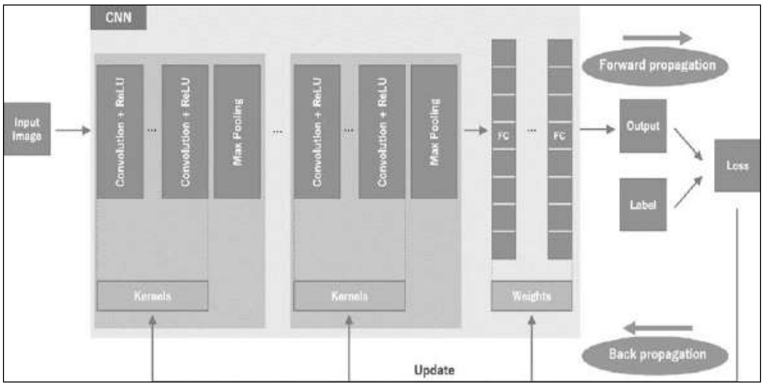

Fig 3: CNN Architecture

Mathematical formulation

System S is represented as

S = {ID, P, F, T, CNN, M}

(1). Input Dataset

ID = {i_1, i_2, i_3…i_n}

Where ID is the input image dataset and i_1, i_2…in are the number of images.

(2). Preprocessing

PR = {pr_1, pr_2, pr_3}

PR stands for preprocessing, and pr1, pr2, and pr3 are the actions that need to be completed during this stage of the process.

- pr1 be the reading of input dataset
- pr2 be the enhancement of image input and
- pr3 be the removal of hair from image.

(3). Feature Extraction

F= {f_1, f_2, f_3…f_n}

Where F is the collection of features that were extracted from the image, and f_1, f_2, f_3, f_n are the characteristics that were extracted, such as the border, thickness, colour, and so on.

(4). Training and Testing file generation

T = {T_1, T_2}

Whereas T is the collection of Training / Testing files, T1 is the Training file, and T2 is the Testing file. Both files contain various values for extracted features, but the Training file also includes the category of each image as either 0 or 1.

(5). Convolutional Neural Network (CNN).

CNN = {C, RL, PO, FC, LS}

Where CNN is algorithm consisting of various stages as

C is convolutional operation

RL be the ReLU activation layer

PO be the Pooling layer

FC be the Full Connection layer and

LS be the Loss function.

(6). Cancer Detection

M= {0, 1}

M is the set of Class having value 0 or 1

0 be the absent of Cancer and

1 be the present of Cancer

4. Result analysis

For the purpose of determining whether or not the training was successful, several metrics are gathered here while the network is being pre-trained. There are a number of different measurements that are taken, including elapsed time, validation accuracy, training precision, and training error. The training accuracy displays the accuracy of classification for each single mini-batch, whereas the validation accuracy displays the accuracy of classification for the dataset as a whole. The loss that occurs in each mini-batch due to the reduction in cross-entropy is referred to as the training loss. Figures 4 & 5 display, all throughout course of seven epochs, the learning parameters for CNN Classification.

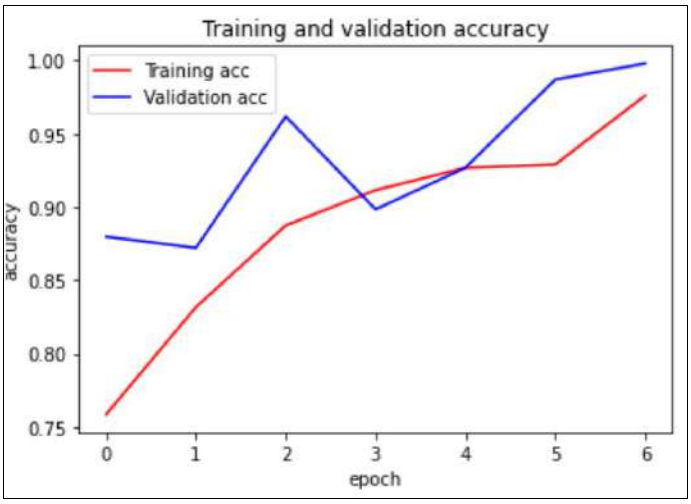

Fig 4: Comparision for Training Accuracy and Validation Accuracy for CNN

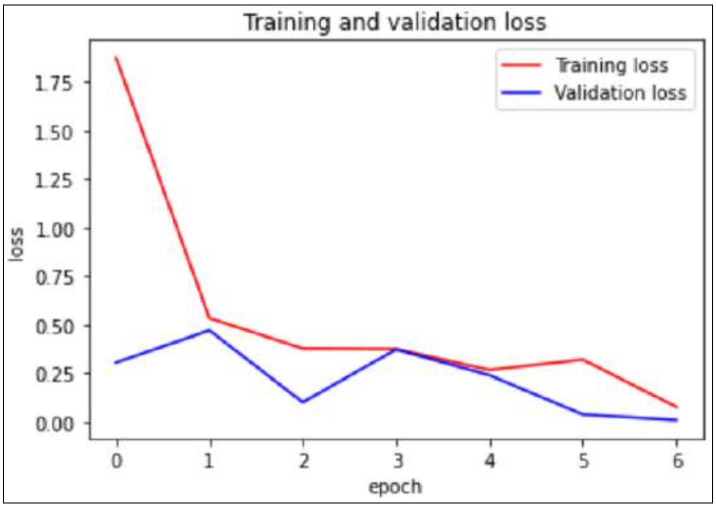

Fig 5: Comparision of Training Loss and Validation Loss for CNN

The results of the trial showed that the CNN algorithm had an accuracy of 99.65 percent overall, with a loss of 0.0067 percent. Figure 6 illustrates the confusion matrix that is used by the CNN algorithm.

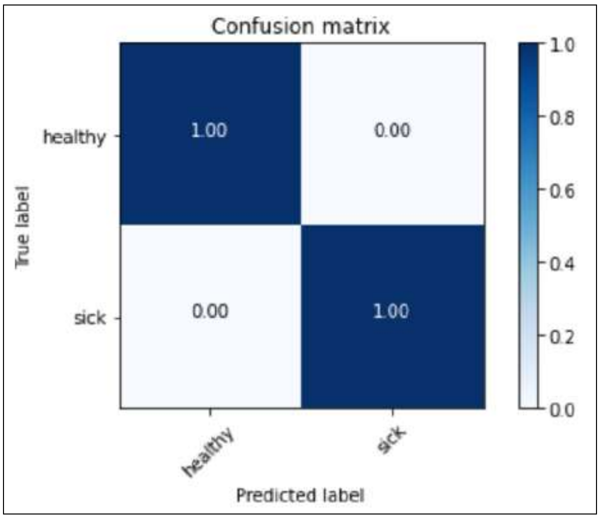

Fig 6: Confusion Matrix for CNN Algorithm

The accuracy of the SVM method came out to be 89.84 percent, whereas the accuracy of the Random Forest algorithm came out to be 90.55 percent when it was calculated on the same dataset. This was done for the purpose of evaluation. Figure 7 presents the results of a comparison of the accuracy of CNN with that of SVM and RF.

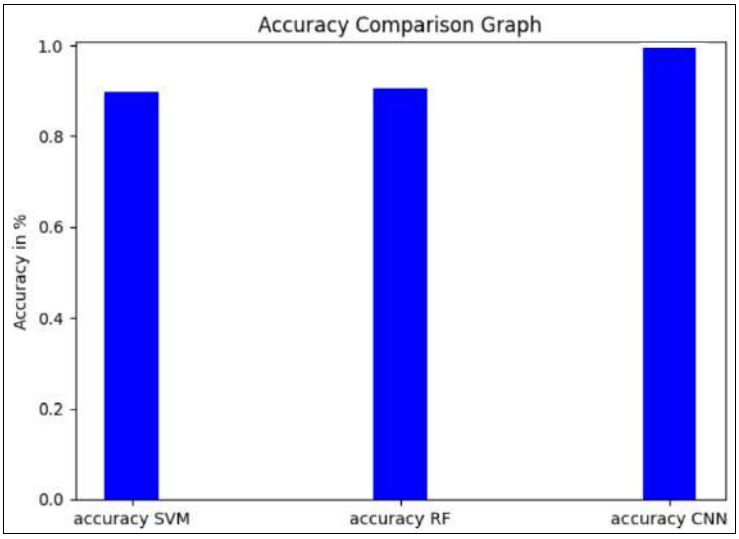

Fig 7: Comparison of Accuracy of CNN, SVM and RF

5. Conclusion

Our research reveals that deep learning techniques that are trained via an end-to-end approach can achieve very high levels of accuracy and have the potential to be easily transferred to a variety of mammography platforms. As the number of available to the public training datasets and computer resources continues to grow, there is a significant possibility that deep learning techniques will be able to considerably improve the effectiveness of breast cancer diagnosis on screening mammography. Within the scope of this paper, we investigated a variety of machine learning approaches for spotting breast cancer. We carried out an investigation into the similarities and differences between CNN, SVM, and Random Forest. It was discovered that CNN performs better than the other approaches that are currently in use in terms of accuracy, precision, and the amount of data that is used. The accuracy that was acquired by CNN was 99.67 percent, whereas the accuracy that was gained by SVM was 89.84 percent, and the accuracy that was obtained by RF was 90.55 percent. Our method may be helpful in the development of more advanced CAD systems in the future. These systems might be used to assist in the prioritisation of the most create a special to be evaluated by a radiologists, or as an automated primary approach once an original impartial interpretation has been made. Our method can also be utilised in the solution of further medical imaging issues, particularly those that include a dearth of ROI annotations.

6. References

1. Zhou X, Li C, Rahaman MM, Yao Y, Ai S, Sun C, *et al*. A Comprehensive Review for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks. IEEE Access. 2020;8:90931-56. https://doi.org/10.1109/ACCESS.2020.2993788.
2. Sutanto DH, Ghani MKA. A Benchmark of Classification Framework for Non-Communicable Disease Prediction : A Review, 2015.
3. Gautam R, Kaur P, Sharma M. A comprehensive review on nature inspired computing algorithms for the diagnosis of chronic disorders in human beings. Prog Artif Intell 2019;8:401-24. https://doi.org/10.1007/s13748-019-00191-1.
4. Waciko KJ, Ismail B. SARIMA-ELM hybrid model versus SARIMA-MLP hybrid model. International Journal of Statistics and Applied Mathematics. 2020;5(2):01-08.
5. Fatima N, Liu L, Hong S, Ahmed H. Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. IEEE Access 2020;8:150360-76. https://doi.org/10.1109/ACCESS.2020.3016715.
6. Wisesty UN, Mengko T, Purwarianti A. Gene mutation detection for breast cancer disease: A review. IOP Conf Ser Mater Sci Eng 2020;830:32051. https://doi.org/10.1088/1757-899X/830/3/032051.
7. Pang T, Wong JHD, Ng WL, Chan CS. Deep learning radiomics in breast cancer with different modalities: Overview and future. Expert Syst Appl 2020;158:113501. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113501.
8. Vasundhara S, Kiranmayee BV, Chalumuru Suresh. Machine Learning Approach for Breast Cancer Prediction, 2019.
9. Muhammet Fatih Ak. A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications, 2020.
10. Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S. Breast Cancer Prediction using Machine Learning, 2019.
11. Ali Raad, Ali Kalakech, Mohammad Ayache. Breast cancer classification using neural network approach; MLP & RBF, 13th International Arab Conference on Information Technology ACIT, ISSN, 2012, 182-0857.
12. Alaa M Elsayad. Predicting the severity of breast masses with ensemble of Bayesian classifiers. Journal of computer science. 2010;6(5);576-584.
13. Miljenko K, Matej Mertik. Application of Bayesian networks in emergency medicine, 2008.
14. Mahmood M, Al-Khateeb B, Alwash WM. A review on neural networks approach on classifying cancers. IAES Int J Artif Intell 2020;9:317-26. https://doi.org/10.11591/ijai.v9.i2.pp317-326.