A Performance Based Study on Deep Learning Algorithms in the Effective Prediction of Breast Cancer

Pronab Ghosh
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
pronab1712@gmail.com

Sami Azam*
College of Engineering, IT and Environment
Charles Darwin University
NT, Australia
sami.azam@cdu.edu.au

Khan Md. Hasib
Department of Computer Science & Engineering
Ahsanullah University of Science & Technology, Dhaka, Bangladesh
khanmdhasib.aust@gmail.com

Asif Karim
College of Engineering, IT and Environment
Charles Darwin University
NT, Australia
asif.karim@cdu.edu.au

Mirjam Jonkman
College of Engineering, IT and Environment
Charles Darwin University
NT, Australia
mirjam.jonkman@cdu.edu.au

Adnan Anwar
School of IT, Centre for the Cyber Security Resaerch and Innovation
Deakin University
Victoria, Australia
adnan.anwar@deakin.edu.au

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9534293

Abstract— Breast Cancer is one of the leading causes of death worldwide. Early detection is very important in increasing survival rates. Intensive research is therefore done to improve early detection of such cancers through the use of available technology. This includes various image processing techniques andgeneral machine learning. However, the reported accuracy for many of these studies was often not at the desirable level. Deep Learning based techniques are a promising approach for the early detection of Breast Cancer. We have therefore done a comparative analysis of seven Deep Learning techniques applied to the Wisconsin Breast Cancer (Diagnostic) Dataset. Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) were proven to be the most effective algorithms as these have demonstrated good results for the majority of performance indicators used in this study, including an accuracy of over 99 percent.

Keywords—breast cancer, deep learning, LSTM, GRU, health informatics, machine learning

I. Introduction

Breast cancer is one of the leading causes of death for women worldwide. The incidence of breast cancer is increasing, especially in developing countries. According to the World Health Organization (WHO), 19.3 million cases of cancer are expected to occur in 2025. Mammography, often in combination with ultrasound, is the most common technique in use for early detection of cancerous breast tumors. MRI is also sometimes used. Generally Mammography is preferred over MRI, as MRI may cause some degree of allergic reactions to some patients. Skilled Radiologists are required to interpret the screening results. Unfortunately, there is a shortage of radiologists around the world, especially in regional areas and underdeveloped countries.

Time is a very important factor in saving lives in the case of breast cancer. However, many countries lack the human resources and technology to deliver prompt patient services for breast cancer screening, diagnosis, and treatment. Image processing techniques along with machine learning could therefore be a valuable weapon in the fight against breast cancer. Several scholars have proposed methods and techniques for detection and diagnosis, but these have often resulted in high false positive and false negative rates. Deep Learning technologies could help to detect breast cancer in an early stage. A number of Deep Learning algorithms are currently available but not all of them have so far been investigated for their usefulness in detecting breast cancer. This research aims to address this issue.

II. Related Work

Studies related to CAD (Computer Aided Detection) for breast cancer focus primarily on the identification and diagnosis of breast tumors. This section briefly outlines the current work relevant to these aspects. For the identification of breast tumors, Sun *et al.* [1] suggested a form of mass detection where an adaptive fuzzy C-means segmentation algorithm is used for each mammogram of the same breast. A supervised artificial neural network is used as a classifier to determine where the segmented zone contains a tumor. This study took advantage of comparison of potential lesions between mammograms of the same breast through means of parallel analysis to enhance the specificity of the CAD results.

Quadri *et al.* [2] suggested that various CAD aproaches are used to verify the consistency of the dataset. They proposed an effective algorithm using a deep learning approach that increases the detection performance based on the WBCD (Wisconsin Breast Cancer Dataset) dataset from the UCI repository. Their presented classifier offers a 99.85% accuracy and positive results relative to previous work by different researchers. The method leads to a new efficient classification paradigm for the diagnosis of breast cancer with high accuracy. Agarap *et al.* [3] presented a comparison of six machine learning (ML) algorithms: Multilayer Perceptron (MLP), GRU-SVM, Linear Regression, Softmax Regression, Nearest Neighbor (NN) Search and Wisconsin Support Vector Machine (SVM) on (WDBC) and calculated the precision of the test classification as well as the sensitivity and specificity values. Their findings indicate that all the ML algorithms performed well (all surpassed 90% test accuracy) for classification. Outstanding among the implemented algorithms is the MLP algorithm with a precision of 99.04%. Zhang *et al.* [4] have researched the features of breast tumours to predict breast cancer. They obtained an F1-score of 93.53% using neural networks. Yap et al. [5] used three distinct deep-learning processes to detect breast cancer in ultrasound images: a patch-based LeNet, a U-Net, and a transfer-learning technique. They used two datasets Dataset A with 306 images (60 malignant and 246 benign), and Dataset B with 163 images (53 malignant and 110 benign). With a True Positive Rate (TPR) of 0.99, a FPR of 0.16 (unchanged), and an F1-score of 0.92, Transfer Learning FCN-AlexNet performed best for Dataset A whereas Transfer Learning yielded the highest True Positive Rate for Dataset B (0.93). The best overall result was a patch-based LeNet with a FPR of 0.09 and an F1-score of 0.91.

Kahn et al. [6] created a Bayesian network to distinguish benign and malignant breast tumors. They did a computer-aided diagnosis of breast cancer using two physical features and fifteen hand-marked probabilistic characteristics. Wang *et al.* [7] used Extreme Learing Machine (ELM) to identify the features of breast tumors and compare them with the SVM classifier. Qiu *et al.* [8] used CNN to forecast the likelyhood of breast cancer by preparing CNN with a vast volume of time-series evidence. Sun *et al.* [9] used deep neural networks to estimate the short-term likelihood of breast cancer, based on 420-time series mammography data. Jiao *et al.* [10] suggested an in-depth feature-based system for breast mass classification, which included CNN and the decision tree methods. Arevalo *et al*. [11] used CNN to abstract breast tumors' images and then identify the cancer as either benign or malignant. Carneiro *et al.* [12] suggested an advanced, deep-learning mammogram prediction approach to predict the risk of breast cancer developing in patients. Kumar *et al.* [13] proposed an image retrieval method using Zernike Moments (ZMs) to retrieve features since characteristics can influence the breast CAD system's efficacy and reliability. Emina *et al.* [14] suggested a breast cancer CAD approach in which genetic algorithms are used to extract meaningful and important characteristics. The Rotation Forest approach has been applied to make decisions for two different types of the CAD, for subjects with or without cancer of the breast. AUC and F measures achieved by the Rotation Forest Classifier and the WBC (Diagnostic) dataset were 99.48%, 99.30% and 99.50% respectively. Wang et al. [15] proposed a mass detection approach based on CNN deep features and Unsupervised Extreme Learning Machine (US-ELM) clustering to create a feature set fusing deep features, morphological features, texture features, and density features. An ELM classifier was then created to identify benign and malignant breast masses using the fused feature set.

III. PROPOSED METHODOLOGY

In our research, various methods are applied for the identification of breast cancer. The aim is to evaluate a range of performance measurement indices for of the above algorithms.

The Wisconsin Breast Cancer (Diagnostic) Dataset (WBDC) from the UCI machine learning repository is used for this research. We analyse seven Deep Learning algorithms, of which some are commonly used while others are used rarely and may need to be investigated further. They are briefly described below.

A. Learning Algorithms

Artificial Neural Network (ANN): Artificial Neural Networks (ANN) are designed using artificial neurons, similar to the neurons of biological brains. The number of such artificial neurons can vary greatly from system to system. They are connected in a series of layers, Input Layer, Hidden Layer and Output Layer. 'Weight' is a unit which determines effect of the connections between these neurons. Weights can both be negative and positive, meaning they either suppress or excite another neuron. Normally information propagates from the Input Layer, through the Hidden Layer(s) to the Output Layer. This is defined as the 'Feedforward' arrangement [18].

'Backpropagation' is the process through which ANNs learn. In this process, the output that has been generated by the network is compared with the one that should have been generated. The difference of these two instances is then taken to adjust the weights between the connections. Over a number of iterations, the network eventually becomes capable of producing a sufficiently accurate result [18].

The working mechanism of ANNs is also the basis of other Deep Learning algorithms.

Recurrent Neural Network (RNN): In a Recurrent Neural Network (RNN), connections between nodes usually generate a directed graph along a temporal sequence, giving it the edge in exhibiting temporal dynamic behaviour [19]. Equation (1) and (2) define how an RNN evolves over time.

$$Q^t = f_1(k^t ; \Phi) \tag{1}$$

$$k^t = f_2(k^{t-1}, j^t ; \Phi) \tag{2}$$

where Q^t is the output of the RNN at time t; j^t is the input to the RNN at time t; and k^t is the state of hidden layers at time t.

As Recurrent Neural Networks work in a feedforward fashion, they can use their internal state, often denoted as ‘memory’, to process inputs sequences that are of variable lengths. This equips RNNs for data processing where data features are seemingly unsegmented but connected in some order [19].

Long Short Time Memory (LSTM): LSTM is an upgraded version of RNN, where it acts as a long short-term memory block within a RNN and is used to establish context for the way the program receives inputs and generates outputs. An LSTM block is an intricate component with various sub-parts such as activation functions, weighted inputs, inputs from previous blocks and the resulting outputs. The unit is named ‘long short-term memory’ block because the program is applying a structure which is based on a short-term memory process creating longer-term memory [20].

Gated Recurrent Unit (GRU): One of the least investigated algorithms especially in works dealing with comparative analysis. GRU is a Gating architecture within RNNs, similar to a LSTM unit except for an output gate. GRU’s are able to handle the issue of Vanishing Gradients which often affects standard RNNs [21]. The Vanishing Gradient problem transpires in machine-based learning models when the gradient becomes vanishingly small, thus preventing the weight from altering its value. A GRU and LSTM both operate on a similar design and often deliver similar results. However, GRU achieves better performance than LSTM, particularly for smaller datasets.

Convolutional Neural Network (CNN): A Convolutional Neural Network (CNN) is a special case of the ANN model and is one of the most commonly used algorithms nowadays [22]. The core of CNN is the *Convolutional Layer* which carries out an operation called a 'convolution'. Convolution is a type of linear operation which involves the multiplication of the inputs with a set of weights. This multiplication is performed between a two-dimensional array of weights, more commonly known as 'filters' and an array of input data. Filters are usually smaller than the input data and help in detecting a suitable feature-map from the input data. CNNs have been proven immensely effective in fields such as Image Processing.

Probabilistic Neural Network (PNN): Probabilistic Neural Networks, less commonly used algorithms in this field, are feedforward neural networks. The algorithm works by approximating the parent probability distribution function (PDF) of each class through a non-parametric function and Kernel Density Estimation. The class probability of new input data is estimated using PDF followed by the application of Bayes’ rule to allocate the class that demonstrates maximal posterior probability to new input data. PNN reduces the probability of misclassification [23]. However, the algorithm has considerably larger memory requirements compared to some other algorithms such as Multilayer Perceptron (MLP) networks. On the other hand, PNNs are relatively insensitive to outliers.

Multi-Layer Perceptrons (MLP): MLP networks are Deep Artificial Neural Networks made of multiple perceptrons. Perceptrons are single layer neural networks [25] between an Input and an Output Layer, MLPs contain an arbitrary number of hidden layers which form the core computational engine of the MLP. Figure 1 shows a general structure of a MLP network.

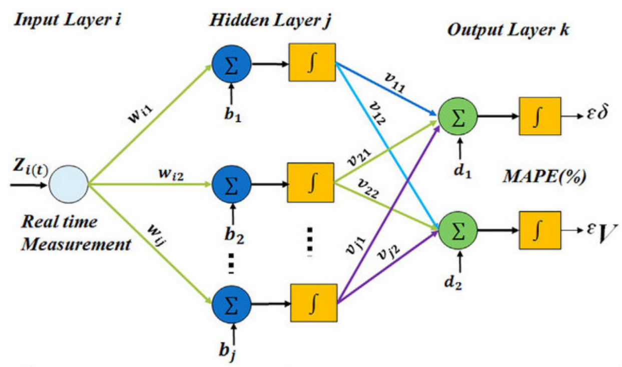

Fig 1. General Structure of a MLP Network [24]

B. Dataset Preprocessing

The Wisconsin Breast Cancer (Diagnostic) Dataset (WBDC) [15] obtained from the UCI machine learning is used in this research. There are a number of research works based on this dataset available in journals and conferences. This dataset has been selected based on the large number of features. Also, it has no missing values.

The values of most attributes of this dataset are in numerical form. An exception is the diagnosis which is categorical. This is transformed into a numeric value to allow processing. Of the 569 instances with 32 characteristics, 357 are of the benign class (B) and 212 are of the malignant class (M). Fig. 2 illustrates this.

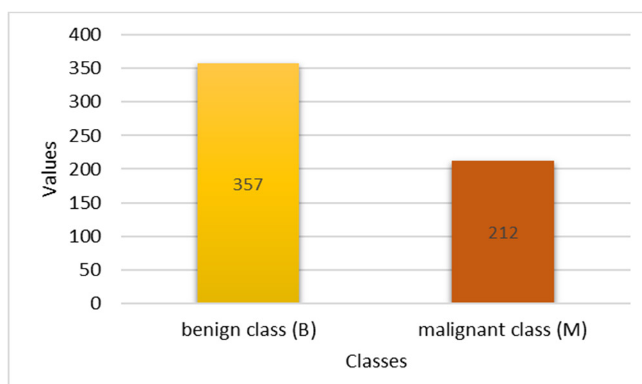

Fig 2 Number of data points of each class.

Benign and Malignant classes are classified as 0 and 1 respectively.

C. Performance Measure Indices

The reliability and consistency of the machine learning system can be measured using success metrics. Positive classification happens when an entity is classified as having malignant. True Positive (TP), True Negative, False Positive (FP) and False Negative (FN) classifications are used to calculate various indices to evaluate the results [16] [17].

TP = True Positive (model correctly identified a breast cancer patient).

TN = True Negative (model correctly identified an individual with no breast cancer).

FP = False Positive (model indicated a non-breast cancer patient as having breast cancer)

FN = False Negative (model failed to identify a patient having breast cancer).

Figure 12 in Appendix A has a detailed flow diagram of the steps taken for performance evaluation.

D. Justification of the Proposed Technique

Algorithms which perform well benefit from a strongly correlated feature set. It is evident from Fig. 3 that the characteristics of radius_mean, texture_mean, perimeter_mean and area_mean have a weak relationship with fractal_dimension_se, symmetry_se, smoothness_se and fractal_dimension_mean with values of about -0.2. On the other hand, a high correlation, close to 0.3, was found for other recorded attributes. Attributes such as texture_se and smoothness_se have little correlation with the features from fractal_dimension_worst to radius_worst. All of the displayed features from radius_worst to symmetry_worst are heavily associated with their linked features with values around 0.3.

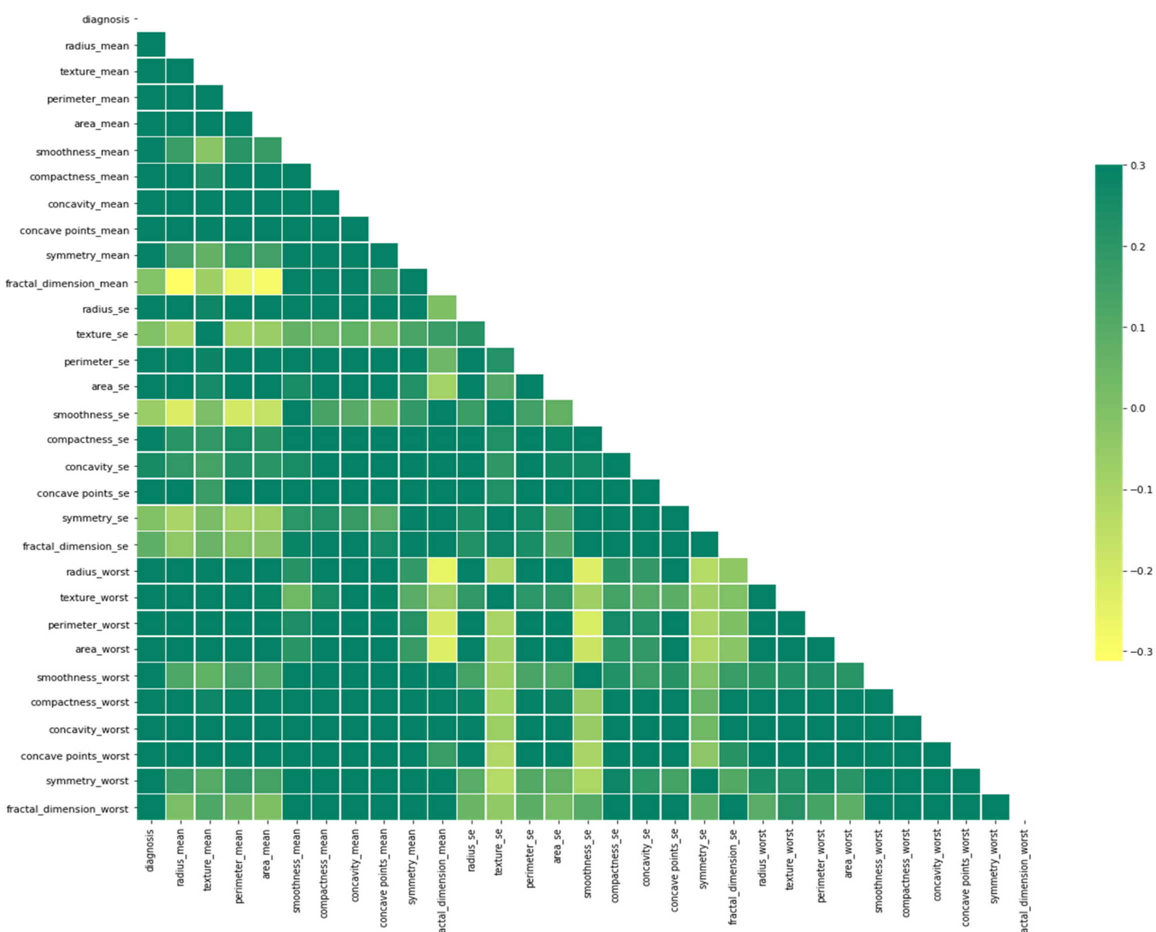

Fig 3. Correlated features of Wisconsin Breast Cancer dataset

IV. RESULTS AND DISCUSSION

A. *Comparison Between Different Methods based on Accuracy*

Accuracy is generally considered to be the most important variable for evaluating the effectiveness of various types of deep learning algorithms. As described above, we are using seven classifiers. Fig 4. shows that GRU and LSTM performed the best with an accuracy of 99.1% each, while RNN achieved the lowest accuracy (approximately 93.9%). The accuracy of the other algorithms varies between 94% and 97.5%.

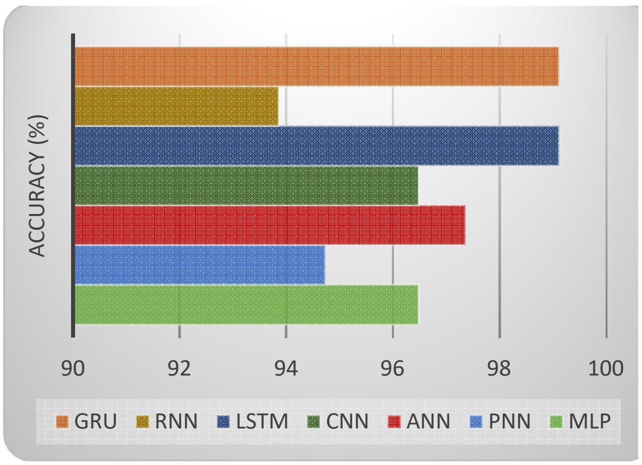

Fig 4. Accuracy

B. *Comparison Between Different Methods based on Sensitivity and Specificity*

Fig. 5 shows the sensitivity and specificity scores of the different algorithms. Five different algorithms, GRU, LSTM, CNN, PNN and MLP, achieved a sensitivity of 100%. RNN and PNN had a lower specificity (about 93%), whereas the GRU and LSTM methods resulted in a sensitivity of just over 98%.

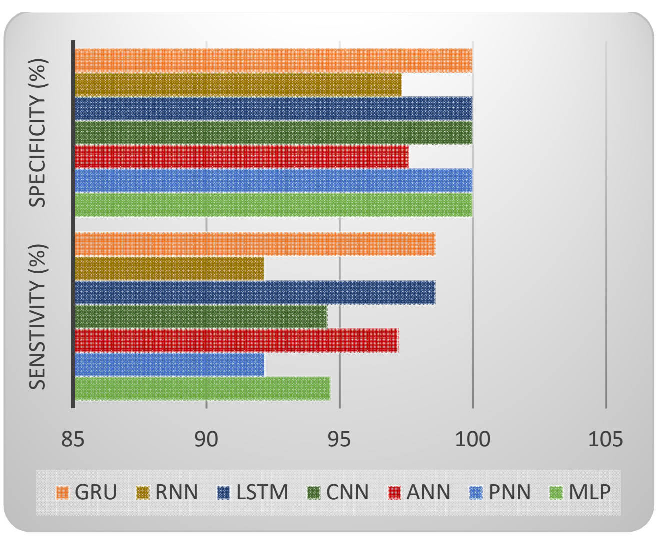

Fig 5. Sensitivity and Specificity

C. *Comparison Between Different Methods based on FPR and FNR*

False Positive Rates and False Negative Rates are commonly considered to be critical in evaluating the effectiveness of diagnostic systems. FPR are zero for all algorithms except RNN which has a FPR of 2.2. RNN also has the highest FNR while the FNR of GRU and LSTM are very low and the FNR of CNN is zero, as shown in Fig. 6.

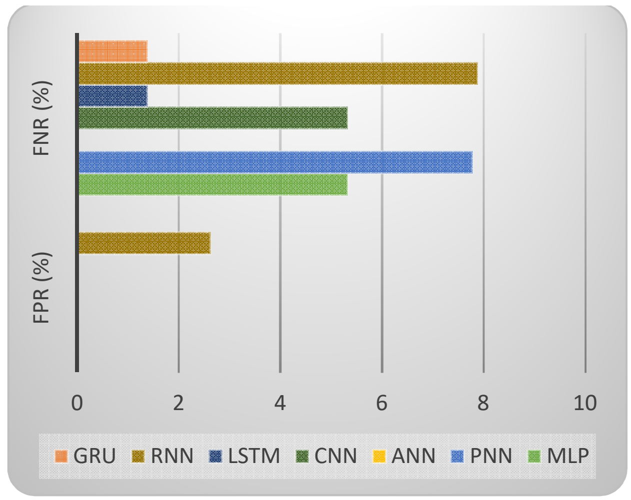

Fig 6. FPR and FNR

D. *Comparison Between Different Methods based on MSE*

The mean squared error (MSE) also applied to evaluate the deep learning algorithms. The lowest error (approximately 0%) is obtained with the ANN approach, while the highest error rate results from the RNN algorithm (just over 6%), see Fig. 7. PNN has the second highest MSE value. CNN and MLP have a MSE of about 3.5% and GRU and LSTM of just under 1%

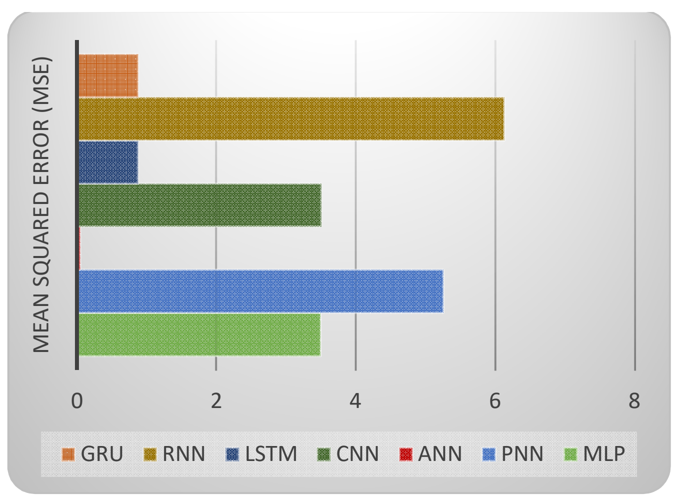

Fig 7. MSE

E. Comparison Between Different Methods based on Log Loss

Figure 8 displays the log loss (LL) for seven forms of deep learning classifiers. The LSTM model produces the lowest LL value of 0.3, while the highest LL value is 2.2 for RNN. Both CNN and MLP have the same LL value of 1.2.

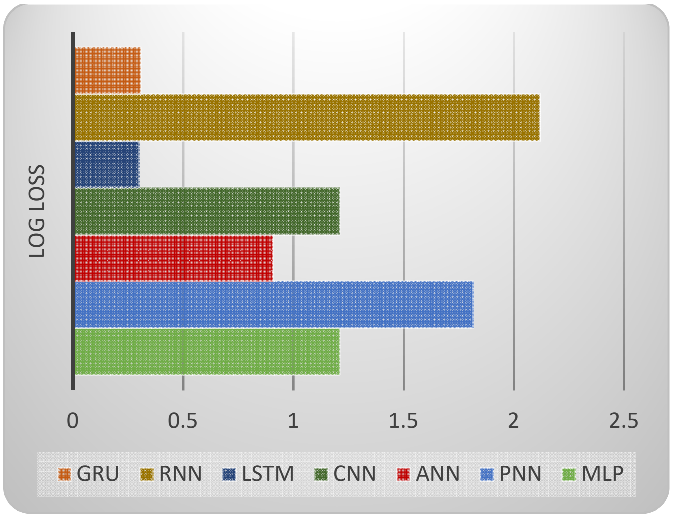

Fig 8. Log Loss

F. Comparison Between Different Methods based on Cohen's Kappa Score

Cohen's kappa coefficient is a metric used to calculate the inter-rater reliability of qualitative products. The highest kappa score is 1 for both GRU and LSTM whereas CNN produces the lowest score of 0.92 as shown in Fig. 9.

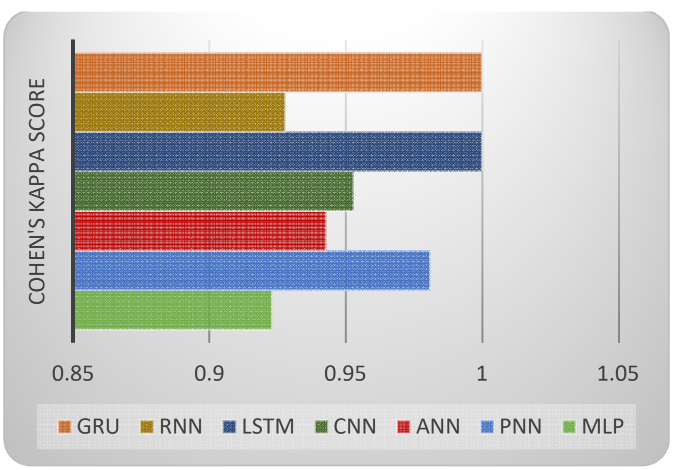

Fig 9. Cohen's Kappa Score

G. Comparison Between Different Deep Learning Methods based on Computation Time

Figure 10 indicates that CNN is the most efficient algorithm in terms of processing time GRU, RNN, PNN and ANN also had reasonable performance with processing times slightly longer than that of CNN. On the other hand both MLP and LSTM required significantly longer processing times.

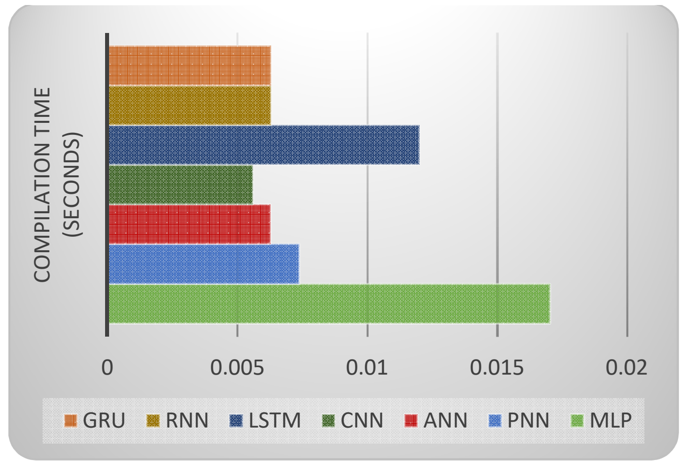

Fig 10. Computation Time

H. Computational Analysis of the Various Approaches

Fig. 11 shows the ranking of the different algorithms for each performance metric. Similar colour indicates same score the algorithms, for instance in terms
of Accuracy, both LSTM and GRU achieved identical scores.

It can be deduced that LSTM performed the best with GRU a close second. RNN was the least successful algorithm.

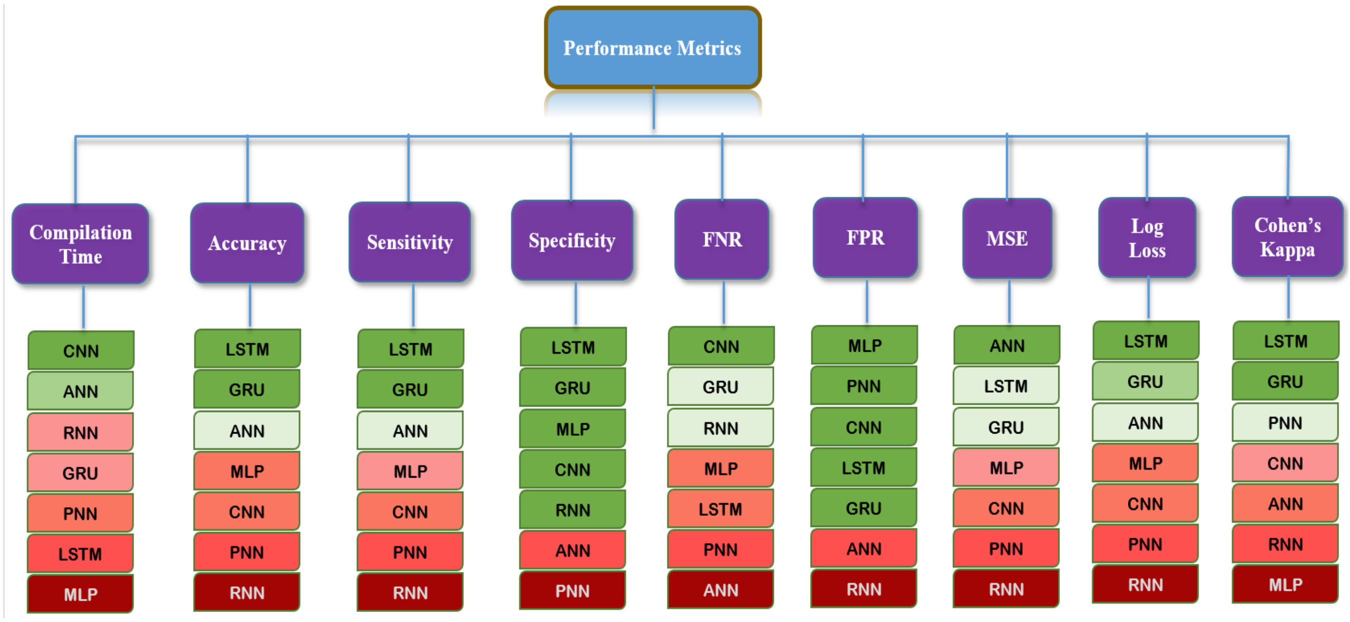

Fig 11. Performance Metrics of several deep learning algorithms

Table 1: Parameters used for the different algorithms

	number of neurons	number of epochs	activation functions	optimiser
ANN	429	128	relu	adam
CNN	3, 346	512	relu	rmsprop
GRU	29851	200	sigmoid	adam
LSTM	293,701	150	sigmoid	adam
MLP	-	100	tanh	adam
PNN	-	-	-	-
RNN	429	200	tanh	adam

V. Conclusions and Future Work

We evaluated seven deep learning models through an array of performance metrics. We also generated a heatmap which contains correlated features and applied various types of optimization tests. From the results above, it is clear that more than one Deep Learning Model can provide classifications in the field of Breast Cancer Detection. Judging from the overall performance metrics, LSTM and GRU can provide useful results. Diagnosis based on Deep Learning based algorithms is a rapidly developing technological arena where new algorithms are proposed regularly. In the future this study will be further expanded to include those upcoming promising algorithms.

Appendix A

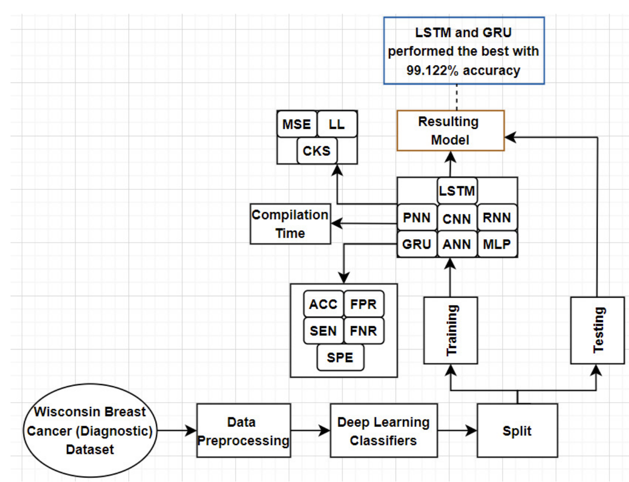

Fig 12. Flow diagram of the evaluations steps

REFERENCES

[1] X. Sun, W. Qian, and D. Song, "Ipsilateral-mammogram computer-aided detection of breast cancer," Computerized Medical Imaging and Graphics the Official Journal of the Computerized Medical Imaging Society, vol. 28, no. 3, pp. 151–158, 2004.

[2] Heidari, Morteza, Abolfazl Zargari Khuzani, Alan B. Hollingsworth, Gopichandh Danala, Seyedehnafiseh Mirniaharikandehei, Yuchen Qiu, Hong Liu, and Bin Zheng. "Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm." Physics in Medicine & Biology 63, no. 3 (2018).

[3] Agarap, Abien Fred M. "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset." In Proceedings of the 2nd intl. conf. on machine learning and soft computing, pp. 5-9. 2018.

[4] Zhang, X., Zhang, Y., Zhang, Q., Ren, Y., Qiu, T., Ma, T., & Sun, Q. (2019). Extracting comprehensive clinical information for breast cancer using deep learning methods. International Journal of Medical Informatics, 132.

[5] M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, and R. Marti, "Automated breast ultrasound lesions detection using convolutional neural networks," IEEE J Biomed Health Inform, vol. 22 (4), 2017.

[6] K. C. Jr, L. M. Roberts, K. A. Shaffer, and P. Haddawy, "Construction of a bayesian network for mammographic diagnosis of breast cancer," Computers in Biology and Medicine, vol. 27, no. 1, pp. 19–29, 1997.

[7] Z. Wang, G. Yu, Y. Kang, Y. Zhao, and Q. Qu, "Breast tumor detection in digital mammography based on extreme learning machine," Neurocomputing, vol. 128, no. 5, pp. 175–184, 2014.

[8] Y. Qiu, Y. Wang, S. Yan, M. Tan, S. Cheng, H. Liu, and B. Zheng, "An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology," in Medical Imaging 2016: Computer-Aided Diagnosis. International Society for Optics and Photonics, 2016.

[9] W. Sun, T. L. Tseng, B. Zheng, and W. Qian, A Preliminary Study on Breast Cancer Risk Analysis Using Deep Neural Network. Springer International Publishing, 2016.

[10] Z. Jiao, X. Gao, Y. Wang, and J. Li, "A deep feature based framework for breast masses classification," Neurocomputing, vol. 197, pp. 221–231, 2016.

[11] J. Arevalo, M. A. G. Lopez, and M. A. G. Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," Computer Methods and Programs in Biomedicine, vol. 127, pp. 248–257, 2016.

[12] G. Carneiro, J. Nascimento, and A. P. Bradley, "Automated analysis of unregistered multi-view mammograms with deep learning," IEEE Transactions on Medical Imaging, vol. 36, no. 11, pp. 2355–2365, 2017.

[13] Y. Kumar, A. Aggarwal, S. Tiwari, and K. Singh, "An efficient and robust approach for biomedical image retrieval using zernike moments," Biomedical Signal Processing and Control, vol. 39, pp. 459–473, 2018.

[14] E. Alickovi ˇ c and A. Subasi, "Breast cancer diagnosis using ga feature selection and rotation forest," Neural Computing and applications, vol. 28, no. 4, pp. 753–763, 2017.

[15] Wang, Z., Li, M., Wang, H., Jiang, H., Yao, Y., Zhang, H., & Xin, J. (2019). Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features. IEEE Access, 7, 105146-105158.

[16] "Breast Cancer Wisconsin (Diagnostic) Data Set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), [Accessed: 03-06-2020].

[17] P. Ghosh, S. Azam, M. Jonkman, A. Karim et al., Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques, in IEEE Access, doi: 10.1109/ACCESS.2021.3053759.

[18] Popel, M. H., Hasib, K. M., Habib, S. A., & Shah, F. M. (2018, December). A hybrid under-sampling method (HUSBoost) to classify imbalanced data. In 2018 21st Intl. Conf. of Computer and Information Technology (ICCIT) (pp. 1-7).

[19] T. Rashid, Make Your Own Neural Network: A Gentle Journey Through the Mathematics of Neural Networks, and Making Your Own Using the Python Computer Language. Charleston, SC, USA: CreateSpace, 2017.

[20] Bang, S., Tijus, C.: Problem Solving using Recurrent Neural Network based on the Effects of Gestures. 10th Intl. Joint Conf. on Computational Intelligence, Spain, 2018.

[21] Y. Fu, S. Sen, J. Reimann, and C. Theurer, "Spatiotemporal Representation Learning with GAN Trained LSTM-LSTM Networks," IEEE Intl. Conf. on Robotics and Automation (ICRA), 2020.

[22] X. Li, X. Ma, F. Xiao, F. Wang, and S. Zhang, Application of Gated Recurrent Unit (GRU) Neural Network for Smart Batch Production Prediction, Energies, vol. 13 (22), 2020.

[23] Mahmud, K., Azam, S., Karim, A., SM, Z., Shanmugam, B., Mathur, D. (2021). Machine Learning Based PV Power Generation Forecasting in Alice Springs. in IEEE Access. Vol. 9, DOI: 10.1109/ACCESS.2021.3066494.

[24] Hasib, K.M., Habib, M. A., Towhid, N. A., & Showrov, M. I. H. A Novel Deep Learning based Sentiment Analysis of Twitter Data for US Airline Service. In 2021 International Conference on Information and Communication Technology for Sustainable Development, 2021.

[25] F. Ancona, A. M. Colla, S. Rovetta et al. Implementing probabilistic Neural Networks. Neural Comput & Applic 5, 152–159 (1997). https://doi.org/10.1007/BF01413860.

[26] Online from: https://www.researchgate.net/publication/314176848_Optimization_of_Neural_Network_Parameters_by_Stochastic_Fractal_Search_for_Dynamic_State_Estimation_under_Communication_Failure/figures?lo=1.

[27] F. M. J. M. Shamrat, Z. Tasnim, P. Ghosh, A. Majumder, M. Z. Hasan "Personalization of Job Circular Advertisement to Candidates Using Decision Tree Classification Algorithm" IEEE Intl. Conf. for Innovation in Technology, 2020.