OPTIMIZED DEEP NEURAL NETWORKS ARCHITECTURE MODEL FOR BREAST CANCER DIAGNOSIS

G. Kanimozhi[1], P. Shanmugavadivu[2]

[1,2] Department of Computer Science and Applications,

[1,2] The Gandhigram Rural Institute (Deemed to be University), Gandhigram, Tamil Nadu, India.

ABSTRACT

Breast cancer has increasingly claimed the lives of women. Oncologists use digital mammograms as a viable source to detect breast cancer and classify it into benign and malignant based on the severity. The performance of the traditional methods on breast cancer detection could not be improved beyond a certain point due to the limitations and scope of computing. Moreover, the constrained scope of image processing techniques in developing automated breast cancer detection systems has motivated the researchers to shift their focus towards Artificial Intelligence based models. The Neural Networks (NN) have exhibited greater scope for the development of automated medical image analysis systems with the highest degree of accuracy. As NN model enables the automated system to understand the feature of problem-solving without being explicitly programmed. The optimization for NN offers an additional payoff on accuracy, computational complexity, and time. As the scope and suitability of optimization methods are data-dependent, the choice of selection of an appropriate optimization method itself is emerging as a prominent domain of research. In this paper, Deep Neural Networks (DNN) with different optimizers and Learning rates were designed for the prediction of breast cancer and its classification. Comparative performance analysis of five distinct first-order gradient-based optimization techniques, namely, Adaptive Gradient (Adagrad), Root Mean Square Propagation (RMSProp), Adaptive Delta (Adadelta), Adaptive Moment Estimation (Adam), and Stochastic Gradient Descent (SGD), is carried out to make predictions on the classification of breast cancer masses. For this purpose, the Mammographic Mass dataset was chosen for experimentation. The parameters determined for experiments were chosen on the number of hidden layers and learning rate along with hyperparameter tuning. The impacts of those optimizers were tested on the NN with One Hidden Layer (NN1HL), DNN with Three Hidden Layers (DNN4HL), and DNN with Eight Hidden Layers (DNN8HL). The experimental results showed that DNN8HL-Adam (DNN8HL-AM) had produced the highest accuracy of 91% among its counterparts. This research endorsed that the incorporation of optimizers in DNN contributes to an increased accuracy and optimized architecture for automated system development using neural networks.

Keywords: *Breast cancer, Neural Networks, Optimization techniques, Gradient Descent*

1. INTRODUCTION:

Cancer is one of the primary causes of mortality and a major barrier in increasing the life span in the entire world. Breast cancer is the major cause of death among all cancer types. The International Agency for Research on Cancer (IARC), a specialized cancer agency for World Health Organization, has reported the cancer incidence and mortality rate updated by Globocan (Global Burden of Cancer Study) in December 2020. As per the report, breast cancer has now overtaken lung cancer and has become the most diagnosed cancer, with 11.7% new cases and 6.9% new death worldwide in 2020 [1,2]. To reduce global breast cancer mortality, early diagnosis with an accurate and reliable procedure is a major concern. Early diagnosis of cancer, prompt access to appropriate treatment and care, palliative and survivorship care, and extensive data collection through robust cancer registries depict an increased survival rate for breast cancer. [3,4]

Mammography is the major screening modality that diminishes breast cancer deaths considerably. This allows physicians to distinguish the breast abnormality between benign (non-cancerous) and malignant (cancerous) cases. Benign tumour doesn't invade neighbouring tissues, which in most cases is harmless. In contradiction, malignant can spread across the body and is extremely dangerous. Researchers have developed a good number of models for the classification of benign and malignant tissue in mammogram images. To enhance the performance of such models, rapid learning is incorporated into images [4]. This approach enables the models to learn from real patient data collected from the routine clinical trials to generate prognostic techniques to guide for future treatment decisions/predictions.

Evidences suggest that about 30% of breast cancers are being missed even by the most experienced radiologists due to the volume of data involved, which costs time-consuming and suffers from inter-radiologist variance. Hence, radiologists realized the need for the techniques and tools that give instant inferences by looking into the patient's medical data to detect cancers and normal cases [5,6]. Deep Neural Networks (DNN) techniques are progressing rapidly for diagnosing diseases with greater accuracy. This approach has an added advantage of performing temporal medical analysis, with a minimum computational cost. Deep Learning (DL), which is a subset of Machine Learning (ML), is designed to stimulate the functionality of the human brain. It shows superior performance in classification problems [7]. The improving effectiveness of DNN approaches is being given a lot of importance by medical practitioners for breast cancer diagnosis. It can predict whether a tumour in a women's breast is malignant or benign. DNN models are expected to perform well, in breast cancer diagnosis, through classification [8].

Optimization plays a vital role in training the neural networks with the suitable set of parameters that can minimize the error rate. With the speed of convergence and the generalization approach, optimization methods can be used with the minimization and maximization functions [9]. Optimization can improve results by helping to choose the inputs that produce the best output.

The contributions of this paper are summarized as follows:

(a) Implemented data preprocessing on missing data and performed one-hot encoding and feature scaling with standardization on the dataset.

(b) Proposed optimized deep neural network architecture model for the automatic classification of benign and malignant cases of cancer.
(c) Applied three deep neural networks model with different optimizers and learning rates to determine the best result in an effective manner.
(d) Compared the performance of the proposed model with various performance metrics for the apposite breast cancer classification and prediction.

The rest of the study is structured in the following manner. Section 2 presents the relevant research works. Section 3 presents the methodology used in this study. Section 4 contains the results and discussions, followed by conclusion and future directions in Section 5.

2. RELATED WORKS:

Many research works on the prediction of breast cancer using Neural Networks (NN) are available literature. This section presents a brief note on the literate on breast cancer detection using the artificial intelligent techniques.

Vijayalakshmi and Mohan Kumar [10] proposed an ensemble classification method for the prediction of breast cancer with the help of feature extraction, Machine Learning Classifier model creation, and Classification. They used Breast Cancer Coimbra Dataset and fusing the Naïve Bayes, Radial basis Function NN and Linear Discriminant Analysis classifiers to predict the presence of breast cancer. The performance of the classifiers are analyzed using accuracy, precision, recall and F1-score. It was observed 75% of accuracy obtained using ensemble method.

Bethapudi et al [11] suggested a feature analysis for breast cancer with genetic-based classification algorithm. They experimented with the algorithm using the 961 instances and 6 salient attributes of Mammographic mass dataset taken from UC Irvine ML repository. Their algorithm implemented the 3-fold cross-validation approach and achieved 84.4% classification accuracy.

Aslan et al [12] undergone the mechanism of analysis of breast cancer on blood analysis data taken from UCIrvine ML repository called BCC dataset. They applied four different ML methods, namely, ANN, SVM, KNN and Extreme Learning Machine (ELM). They used hyperparameter optimization method to get the best accuracy values and obtained 80% accuracy with less training time while using ELM method.

Usha rani [13] proposed a parallel approach by using NN technique for the diagnosis of breast cancer. The dataset used for this technique was obtained from the University of Wisconsin hospital. The dataset was trained with single layer and multilayer models on feed-forward NN, backpropagation learning algorithm with momentum, and variable learning rate. The performance was produced and observed with 92% of accuracy on multilayer NN model.

Kusuma et al [14] proposed a Backpropagation NN optimization method using Nelder Mead for classifying the breast cancer appearance. Two different datasets were used namely, BCC dataset and Wisconsin Breast Cancer Dataset. Moreover, 10-fold cross validation is applied on both datasets and obtained 73% and 89.8% accuracy on the respective datasets.

Dogo et al [15] performed a comparative analysis of gradient-based optimization algorithms on three different image datasets. They used to train the model with optimization techniques such as Adaptive Gradient (Adagrad), Root Mean Square Propagation (RMSProp), Adaptive

Delta (Adadelta), Adaptive Moment Estimation (Adam), Nesterov - accelerated Adaptive Moment Estimation (Nadam) and Stochastic Gradient Descent (SGD). The performance of the optimizers was obtained by evaluating its convergence time, accuracy and loss. Each dataset was produced different results with different accuracies with the convergence time.

3. METHODOLOGY:

EXPERIMENTAL DATASET:

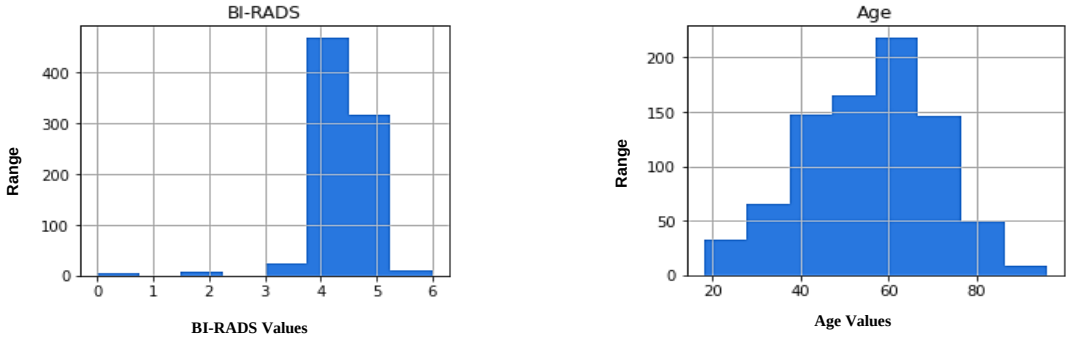
The analysis of optimization algorithms on breast cancer data used Mammographic Mass dataset obtained from UC Irvine Machine Learning Repository [16]. It has 961 instances and 6 attributes in total. The proposed work was conducted based on the clinical data features of the patients. The detail of the dataset is described in Table 1.

Table 1: Description of Mammographic Mass Dataset

Attribute	Type	Description
BI-RADS	Integer	Definitely Benign (1) to highly suggestive of malignancy (5)
Age	Integer	Patient's age in years (18 - 96 years)
Shape	Integer	Mass shape (1-Round, 2-Oval, 3-Lobular, 4-Irregular)
Margin	Integer	Mass margin (1-Circumscribed, 2-Microlobulated, 3-Obscured, 4-Ill-defined, 5-Spiculated)
Density	Integer	Mass density (1-High, 2-ISO, 3-Low, 4-Fat-containing)
Severity	Integer	Predictor class (0-Benign; 1-Malignant)

Breast masses on mammogram are described in terms of shape, margin and density. Margins are the most reliable indicator of the possibility of malignancy, circumscribed margins are the best predictor of a benign lesion, and speculated margins are highly suspicious for malignancy [17].

Figure 1 showed the frequency of data distributed in the mammographic mass dataset.



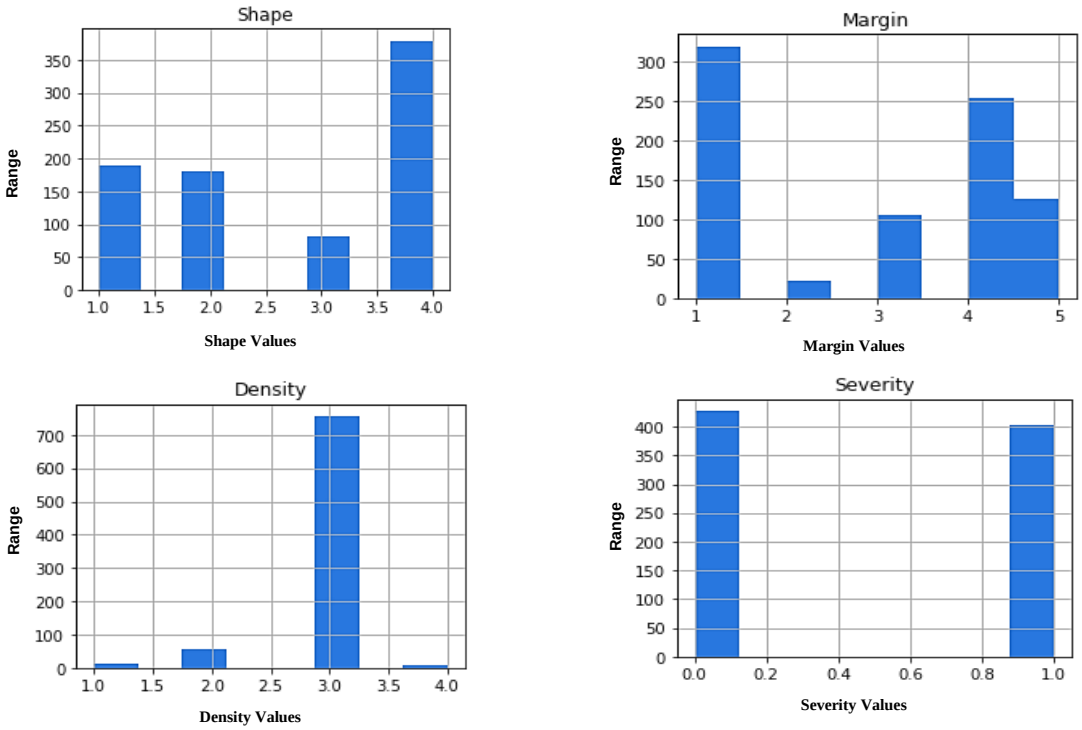


Figure 1: Frequency plots of the attributes

From the figure 1 plots, it is concluded that all the features contained a small amount of missing values. Removing these missing values may not affect the performance of the dataset, since it is very low in number.

DATA PREPROCESSING:

Preprocessing the data is an important task to get better performance because data in the dataset may have noisy, missing values of false data. Data preparation, cleaning, and transformation comprises a major role to make data suitable and impact the accuracy of the model. One-hot encoding, also known as binary encoding technique, is one of the data encoding techniques during preprocessing phase. It is used to convert the categorical or text data into binary format which is 0's and 1's [18]. According to the dataset used in this work, One-hot encoding is performed on the four attributes (BIRADS, shape, margin, density). After one-hot encoding, the attribute's data contained values of atmost 0's and 1's, may cause inefficient learning on the patterns during training phase.

Feature scaling is another imporant preprocessing method to scale down the data in order to make the data in equal range. For this purpose, standardization is used to replace the values by their z-scores [19] and is given in Eqn 1.

$$X_{stand} = \frac{X - \mu(X)}{\sigma(X)} \qquad (1)$$

which means, it redistributes the features with their mean $\mu = 0$ and standard deviation $\sigma = 1$.

DEEP NEURAL NETWORKS:

DNN is a powerful machine learning architecture to learn arbitrary input / output functions given through training data. It is the functional unit of Deep Learning and derived from the

way human brain works to solve complex data-driven problems. DNN is defined as training a NN with many hidden layers. DNN has one input layer followed by several hidden layers and one output layer. The data is entered in the input layer and it is forwarded to the hidden layers and finally the result is passed to the output layer [20]. In this work, three types of NN models were used: NN1HL (Neural Networks with One Hidden Layer), DNN4HL (Deep Neural Networks with Four Hidden Layers), and DNN8HL (Deep Neural Networks with Eight Hidden Layers) additionally with dropout layer. ReLu is used as an activation function on all the models and sigmoid is used as output activation function. Figure 2(a) and 2(b) illustrates the DNN4HL model and DNN8HL model respectively.

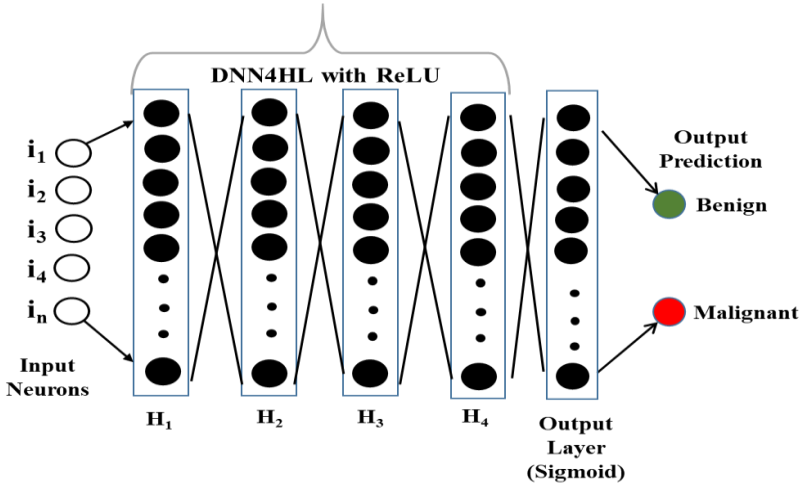


Figure 2(a): DNN4HL Model

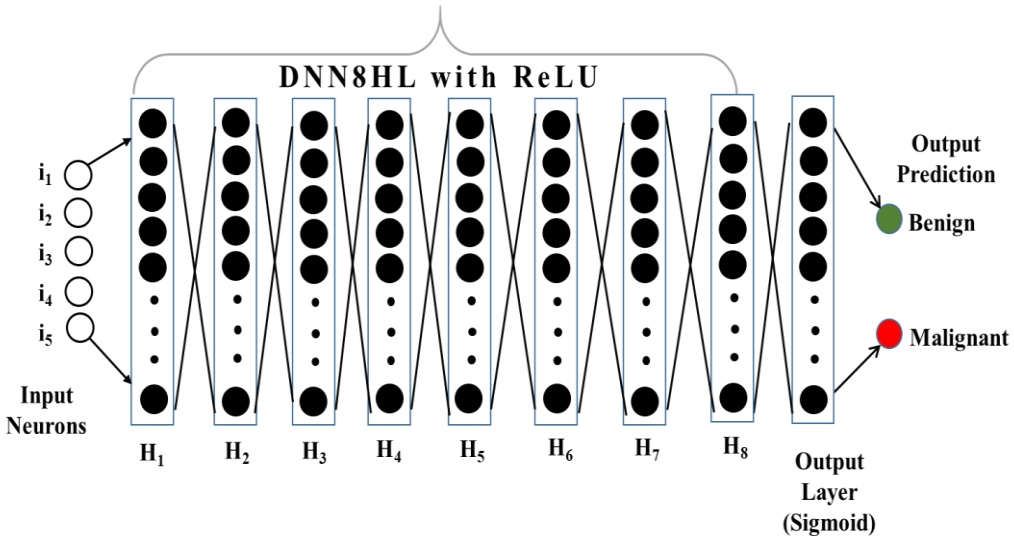


Figure 2 (b): DNN8HL Model

OPTIMIZATION TECHNIQUES:

Optimization is a technique for identifying input parameters or arguments that minimize the cost function of the weight parameter of the trained data , resulting in fewer errors on the test data. The key objective of optimization is to reduce the training error and the gap between training errors and testing errors [21]. Gradient Descent is an optimization algorithm which is inherited as a multivariate continuous objective function and it is a first-order algorithm for

optimizing the target objective function as it explicitly uses a first-order component. A slope or rate of change of an objective function at a given point is called the first-order component. The procedure entails first calculating the function's gradient, then following it with a step size in the opposite direction (for example, downward to the minimum for minimization issues) (also called the learning rate). Each iteration of the algorithm is controlled by the learning rate, which is a hyperparameter that governs the movement against the gradient. Learning rate is a floating point value that makes the proportion of weights updated. Larger values result in faster early learning, before the rate is updated whereas the smaller values lead to slow learning during training [22].

VARIANTS OF GRADIENT DESCENT ALGORITHMS:

In this work, the first order component method involves the following variants of gradient descent algorithms, namely, Adam, RMSProp, Adagrad, Adadelta, and Stochastic Gradient Descent.

The configuration parameters [23] used in optimizers are:

- **learning rate or step size:** The proportion (floating point value) that weights are updated.
- **epsilon:** Very small number to prevent any division by zero in the implementation (default value:10e-8).
- **rho:** Discounting factor for the history/coming gradient. (default value:0.9)
- **beta1:** The exponential decay rate for the first moment estimates (default value: 0.9).
- **beta2:** The exponential decay rate for the second-moment estimates (default value: 0.999).
- **momentum:** float hyperparameter >= 0 that accelerates gradient descent in the relevant direction (default value:0)
- **nesterov:** boolean, whether to apply nesterov momentum (default value: false)

ADAGRAD:

Adagrad is a parameter-specific optimizer for learning rates by using cumulative sum of squared gradients. This is faster for parameters with larger gradients and slower for smaller gradients that results in reducing the learning rate. The hyperparameters used are learning rate and epsilon [24].

RMSPROP:

Instead of using cumulative or accumulative sum of squared gradients like Adagrad, RMSProp uses exponentially decaying average of squared gradient and does not consider the history from the extreme past. As a result, the algorithm converges rapidly once it finds the locally convex bowl [25]. The hyperparameters used are learning rate, rho, and epsilon.

ADAM:

Adam refers to Adaptive moment estimation that computes adaptive learning rates for each parameter. It is an optimization algorithm that can be used to update the network weights iteratively using training data rather than the classical stochastic gradient descent procedure [26]. In order to converge faster, Adam uses momentum and adaptive learning rates. The hyperparameters used are learning rate, beta1, beta2, and epsilon.

SGD:

SGD is similar to the gradient descent algorithm. Additionally, it has momentum, which takes only one sample or small subset of samples at each step for training. It is smarter while used for large datasets with lot of parameters, and can easily update the parameters when new data arrives [27]. The hyperparameters used are learning rate, momentum, and nesterov.

ADADELTA:

Adadelta optimization is a stochastic gradient descent method based on adaptive learning rate per dimension that addresses two issues: the continuous decay of learning rates during training and the requirement for a manually selected global learning rate. Instead of accumulating all prior gradients, it is a more robust variant of Adagrad that adapts learning rates based on a moving window of gradient updates [28]. The hyperparameters used are learning, rho, epsilon.

Figure 3 depicts the overall workflow of the proposed Optimized Deep Neural Networks Architecture Model (ODAM).

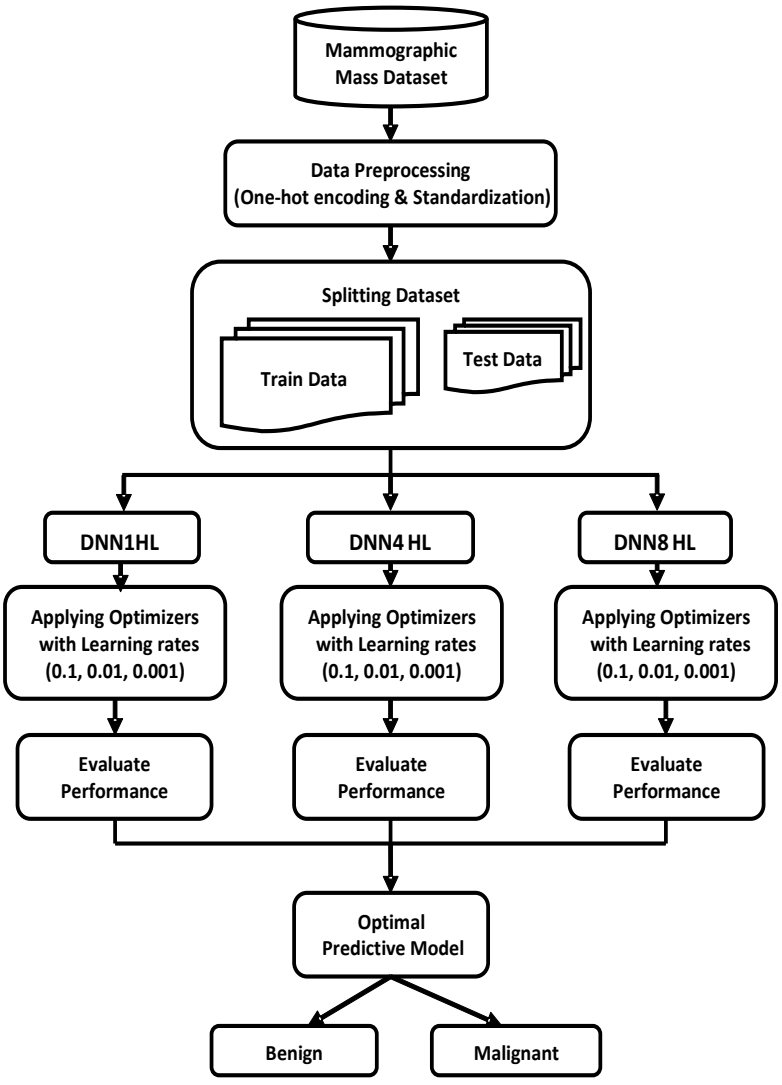


Figure 3: Workflow of Proposed ODAM

The algorithmic description of the proposed ODAM is shown in Algorithm 1.

Algorithm 1: Description of ODAM
Input: Mammographic Mass dataset Output: Prediction of benign or malignant
Begin
Read the dataset Pre-processing: One-hot encoding and Standardization Split the dataset: train/test set Build the networks: DNN1HL, DNN4HL, and DNN8HL with optimizers and learning rates Apply Optimizers (Adam, RMSProp, Adagrad, Adadelta, and SGD) and learning rates (0.1, 0.01, 0.001) Measure the performance: prediction accuracy on the testing set Calculate evaluation metrics: accuracy, confusion matrix, and receiver operating characteristic (ROC) curve
End

4. RESULTS AND DISCUSSION:

This proposed work used three types of network models with different hidden layers and five optimizers (Adam, RMSProp, Adagrad, Adadelta, and SGD) with varied learning rates. Initially, the mammographic mass dataset which has 5 input features, is loaded and preprocessed with one-hot encoding and data standardization techniques. The number of features are now transformed into 21 features. Then the dataset is sent for train/test splitting as the ratio of 75:25 respectively. The splitted data are sequentially modeled as NN1HL with one input layer, 1 hidden layer with Relu activation and one output layer sigmoid activation function. After compiling the model, the parameters are passed through the optimizers with learning rates as 0.1, 0.01, and 0.001 respectively. For testing the data, the model is fitted by using batch size, epoch as parameters and the performance of the model is evaluated. The same procedure is followed for DNN4HL and DNN8HL, with the difference in hidden layers and learning rate as 0.1, 0.01, 0.001 respectively.

Figure 4 illustrates the performance of the optimizers for different learning rates on NN1HL model.

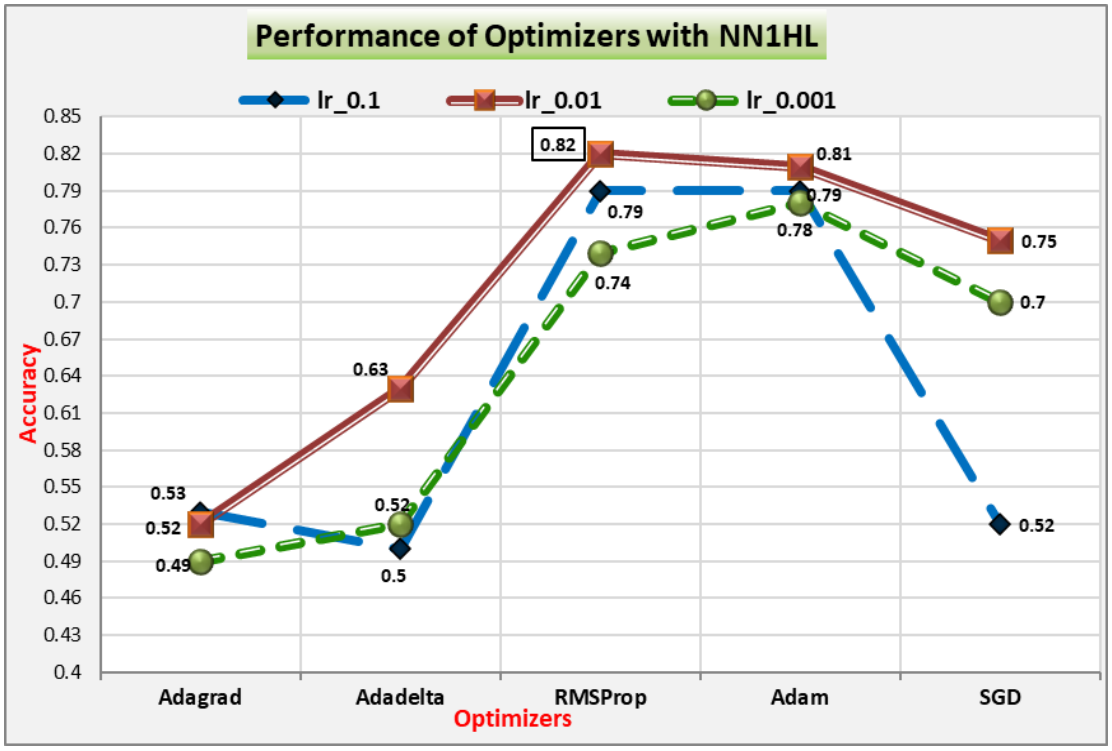


Figure 4: Performance of optimizers on NN1HL Vs. learning rates

From figure 4, it is obvious that the RMSProp optimizer with learning rate 0.01 has a good result with 82%. It is found that the learning rate when used 0.01 has given good performance than the other learning rates on most optimizers. Figure 5 depicts the performance of the optimizers for different learning rates on DNN4HL model.

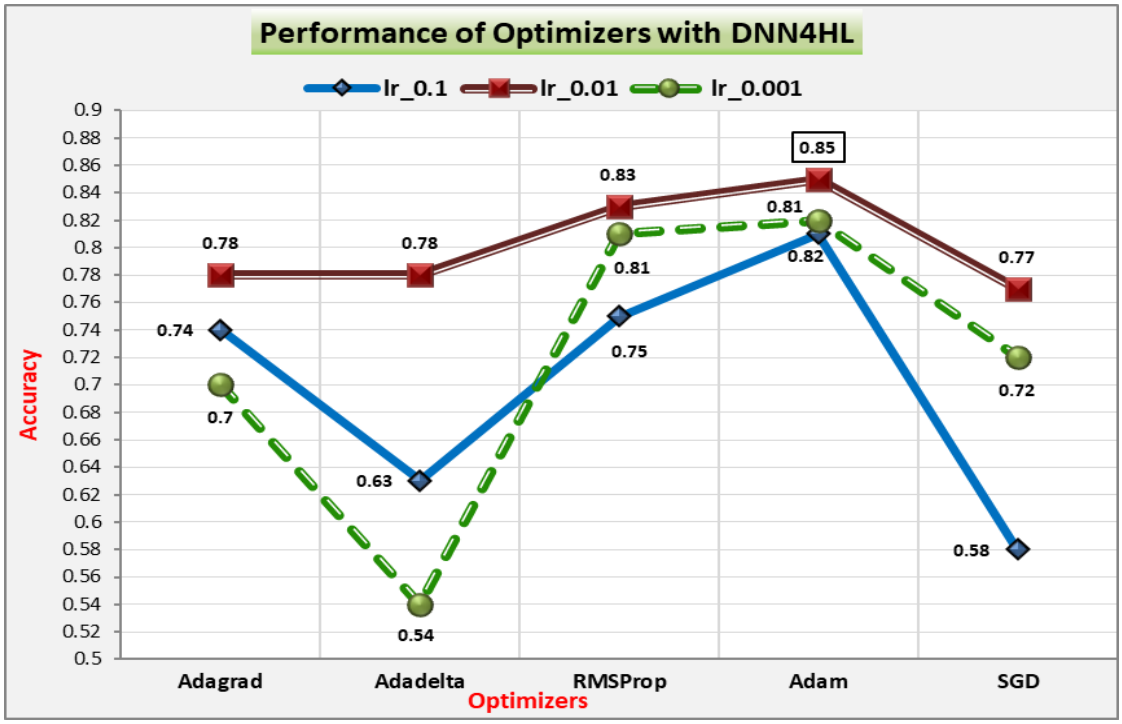


Figure 5: Performance of optimizers on DNN4HL Vs. learning rates

From figure 5, it is found that DNN4HL exhibits good performance on Adam for the learning rate of 0.01. It is also found that the learning rate when used 0.01 has given good accuracy than

the other learning rates on all the optimizers Though this model produced good results, it is insufficient for a dataset to predict its classification accuracy. Figure 6 illustrates the performance of the optimizers for different learning rates on DNN8HL model.

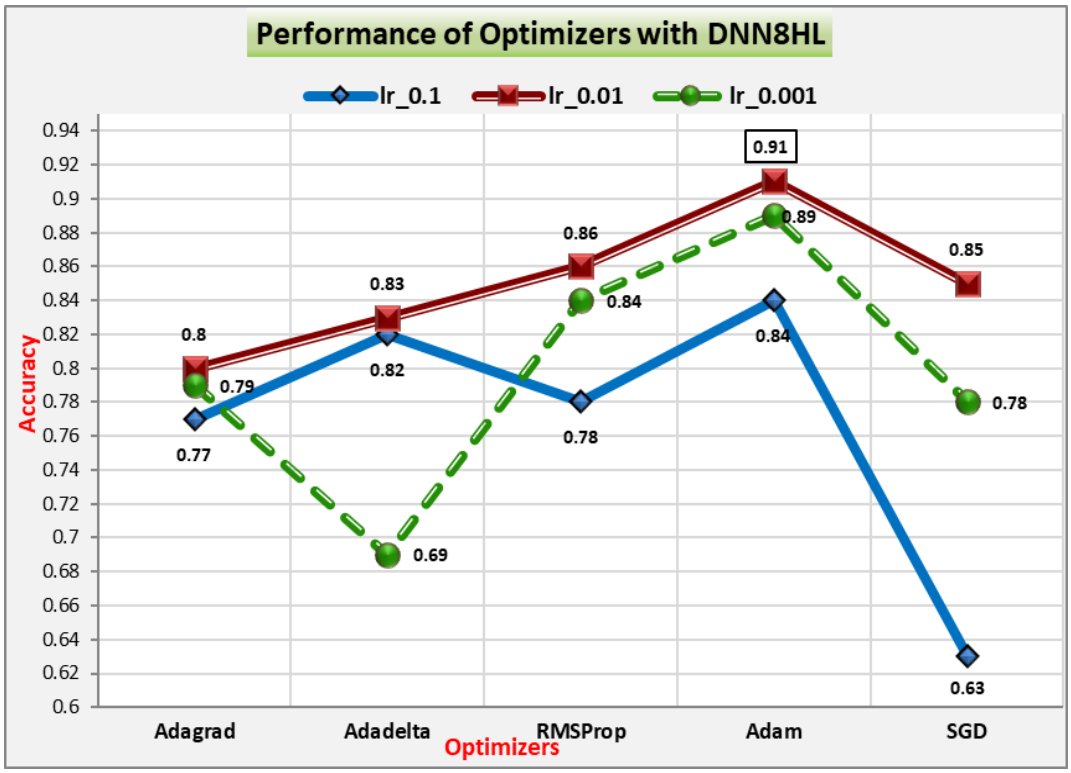


Figure 6: Performance of Optimizers on DNN8HL Vs. learning rates

The figure 6 clearly depicts that the performance of the DNN8HL model with Adam optimizer produced higher accuracy when compared with all the other optimizers on learning rate 0.01. Since the convergence of Adam is faster than other variants because it is the combination of momentum and RMSProp. Thus, the results produced by figure 6 clearly indicates the good level of performance of Adam optimizer.

Table 2 reveals the performance of all the three models, NN1HL, DNN4HL, and DNN8HL while applied different optimizers and learning rate. Figure 7 shows the comparative analysis of optimizers with respect to the DNN models.

Table 2: Performance of Optimizers on DNN models for different learning rates

Models	**NN1HL**			**DNN4HL**			**DNN8HL**		
Optimizers	**lr_0.1**	**lr_0.01**	**lr_0.001**	**lr_0.1**	**lr_0.01**	**lr_0.001**	**lr_0.1**	**lr_0.01**	**lr_0.001**
Adagrad	0.53	0.52	0.49	0.74	0.78	0.7	0.77	0.8	0.79
Adadelta	0.5	0.63	0.52	0.63	0.78	0.54	0.82	0.83	0.69
RMSProp	**0.79**	**0.82**	0.74	0.75	0.83	0.81	0.78	0.86	0.84
Adam	**0.79**	0.81	**0.78**	**0.81**	**0.85**	**0.82**	**0.84**	**0.91**	**0.89**
SGD	0.52	0.75	0.7	0.58	0.77	0.72	0.63	0.85	0.78

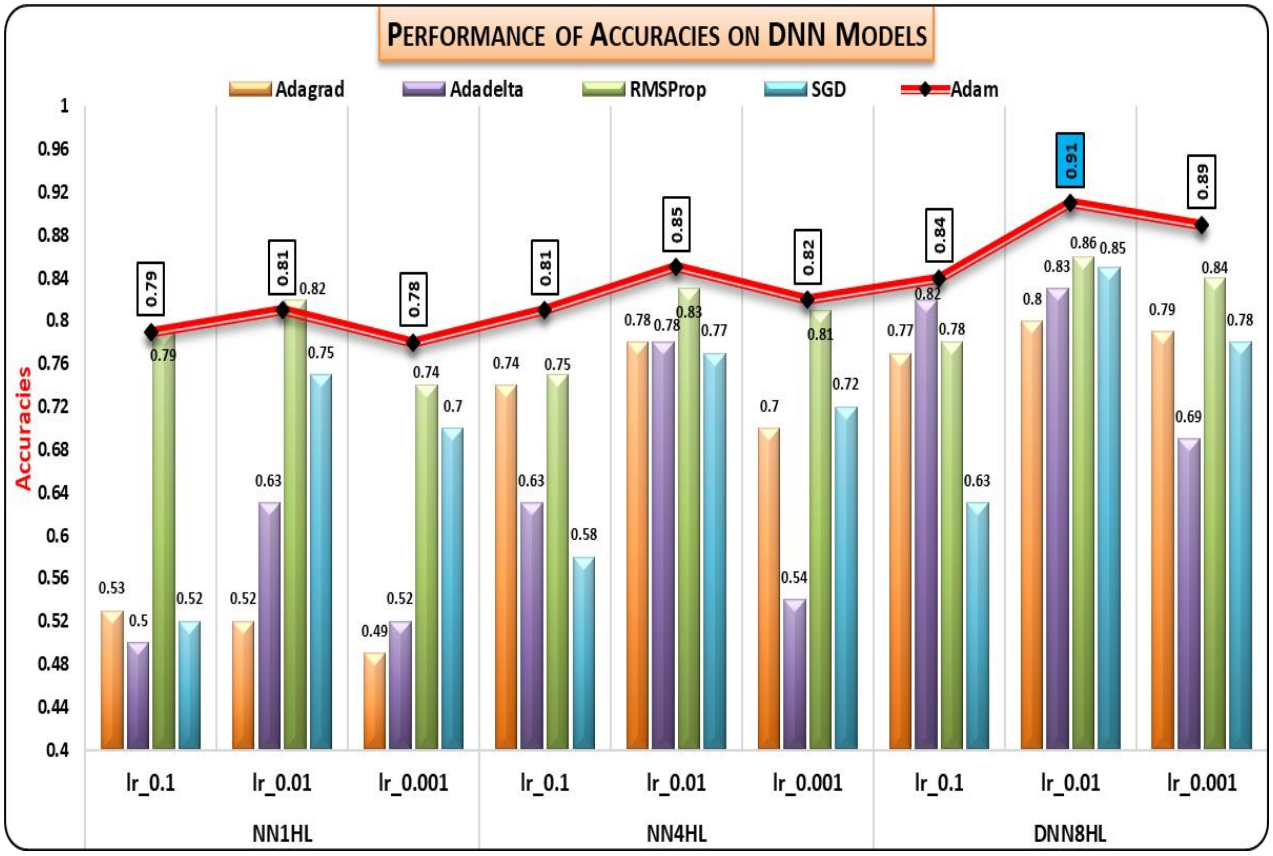


Figure 7: Performance of DNN Models

The figure 7 illustrates the accuracy of all the DNN models used and comparison is done with all the models on different optimizers ad learning rates with respect to its accuracy. In addition to that, Figure 7 depicts that among the five optimizers, Adam has the highest accuracy with 91% on 0.01 as learning rate and DNN8HL-Adam (DNN8HL-AM) optimizer outperformed well on all the three models and it shows consistent increase in the performance accuracy against other optimizers.

The confusion matrix of the proposed optimal DNN8HL-AM on mammographic mass test data is depicted in figure 8. It represents the classification report based on the test set of the proposed approach of mammographic mass dataset.

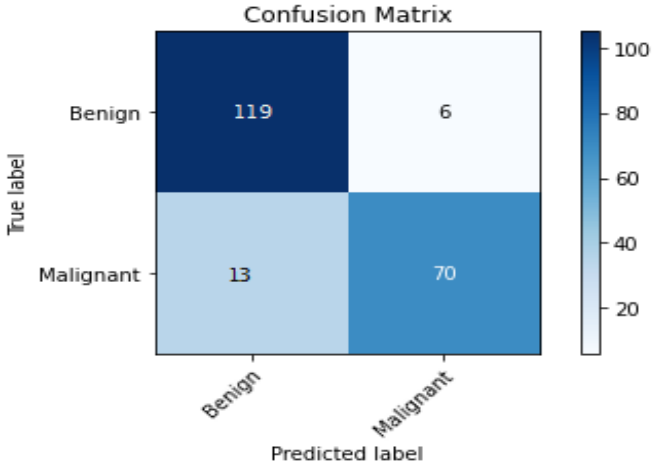


Figure 8: Confusion Matrix for DNN8HL-AM on 0.01 learning rate

From the figure 8, it is observed that out of 208 testing cases, 119 cases were classified, and predicted as benign cases which is true positive (TP). 70 cases were observed and predicted as malignant cases which are true negative (TN). 13 cases were actually identified as benign but predicted as malignant cases which were described as false negative (FN). 6 cases were observed as malignant, but predicted as benign cases which represented as false positive (FP). According to the confusion matrix report, the accuracy rate is observed as 90.8% on the test cases and 9.1% is observed as the error rate.

The ROC curve of the proposed DNN8HL-AM by plotting the TP rate against the FP rate is shown in figure 9.

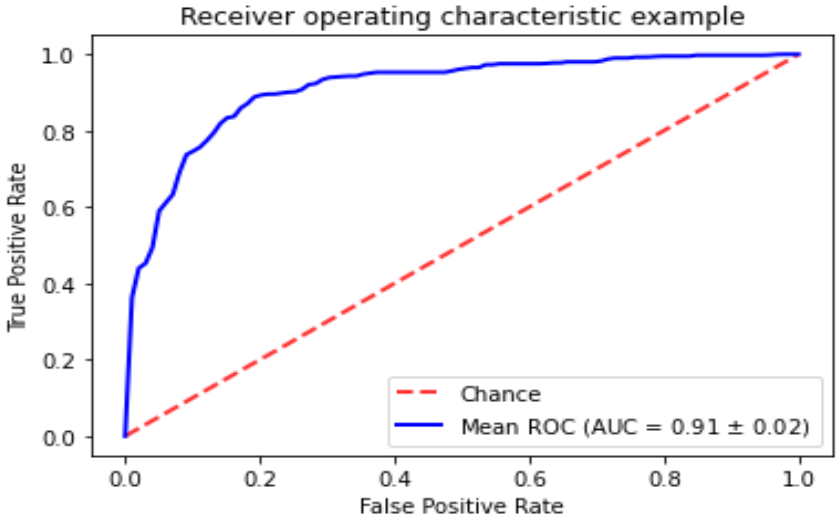


Figure 9: ROC curve of DNN8HL-AM

The figure 9 summarized that the performance of the proposed optimized DNN8HL-AM shows a high AUC value of 0.91 ± 0.02. It represents that, with 91% chance, the proposed model will be able to distinguish between benign and malignant cases. According to outcomes achieved, the proposed method has achieved a better classification accuracy.

5. CONCLUSION:

This paper presented an experimentation on gradient descent optimizers performed on DNN. It is analyzed with five different types of optimizers on three types of models. The experimental dataset used for this proposed analysis was Mammographic mass clinical dataset. The major role of this paper is to investigate the performance of the DNN on breast cancer prediction. The results demonstrated that Adam optimizer with the learning rate as 0.01 has given the highest performance on the breast cancer dataset which suggests its application on the other datasets for better classification and prediction. This research work has a future scope to develop a novel optimizer with optimal update rules for robust prediction rate and accuracy in breast cancer prognosis. This can also be further augmented with image dataset to complement the robust outcomes of research.

ACKNOWLEDGMENTS

The authors record their acknowledgements for the physical computing facilities used for their experimental study and the financial assistance funded under the UGC-SAP (DRS-I) project.

REFERENCES

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

[2]. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3), 209-249.

[3]. Mertz, S., Mayer, M., Paonessa, D., Papadopoulos, E., Alessandro, F., Peccatori, K. S., ... & Spence, D. (2016). Breast Cancer Center Survey: Cancer center management, support, and perception of mBC patient needs across 582 healthcare professionals.

[4]. Prager, G. W., Braga, S., Bystricky, B., Qvortrup, C., Criscitiello, C., Esin, E., ... & Ilbawi, A. (2018). Global cancer control: responding to the growing burden, rising costs and inequalities in access. ESMO open, 3(2), e000285.

[5]. Suleiman, W. I., Rawashdeh, M. A., Lewis, S. J., McEntee, M. F., Lee, W., Tapia, K., & Brennan, P. C. (2016). Impact of Breast Reader Assessment Strategy on mammographic radiologists' test reading performance. *Journal of medical imaging and radiation oncology*, *60*(3), 352-358.

[6]. Mittal, D., Gaurav, D., & Roy, S. S. (2015, July). An effective hybridized classifier for breast cancer diagnosis. In 2015 IEEE international conference on advanced intelligent mechatronics (AIM) (pp. 1026-1031). IEEE.

[7]. Bini, S. A. (2018). Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care?. *The Journal of arthroplasty*, *33*(8), 2358-2361.

[8]. Karthik, S., Perumal, R. S., & Mouli, P. C. (2018). Breast cancer classification using deep neural networks. In *Knowledge computing and its applications* (pp. 227-241). Springer, Singapore.

[9]. Sun, S., Cao, Z., Zhu, H., & Zhao, J. (2019). A survey of optimization methods from a machine learning perspective. *IEEE transactions on cybernetics*, *50*(8), 3668-3681.

[10]. Mahesh, V. G., & Kumar, M. M. (2021, February). An ensemble classification based approach for breast cancer prediction. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1065, No. 1, p. 012049). IOP Publishing.

[11]. Bethapudi, P., Reddy, E. S., Sitamahalakshmi, T., &Varma, K. V. Feature Analysis and Classification of BI-RADS Breast Cancer Using Genetic Algorithm.

[12]. Aslan, M. F., Celik, Y., Sabanci, K., &Durdu, A. (2018). Breast cancer diagnosis by different machine learning methods using blood analysis data. *International Journal of Intelligent Systems and Applications in Engineering*, *6*(4), 289-293.

[13]. Rani, K. U. (2010). Parallel approach for diagnosis of breast cancer using neural network technique. *International Journal of Computer Applications*, *10*(3), 1-5.

[14]. Kusuma, E. J., Shidik, G. F., &Pramunendar, R. A. (2020). Optimization of Neural Network using Nelder Mead in Breast Cancer Classification. *International Journal of Intelligent Engineering and Systems*, *13*(6), 330-337.

[15]. Dogo, E. M., Afolabi, O. J., Nwulu, N. I., Twala, B., &Aigbavboa, C. O. (2018, December). A comparative analysis of gradient descent-based optimization algorithms on convolutional

neural networks. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)* (pp. 92-99). IEEE.

[16]. Mammographic Mass Dataset available at UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

[17]. Elter, M., Schulz-Wendtland, R., & Wittenberg, T. (2007). The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical physics*, *34*(11), 4164-4172.

[18]. Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, *175*(4), 7-9.

[19]. Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. (2014). Data normalization and standardization: a technical report. *Mach Learn Tech Rep*, *1*(1), 1-6.

[20]. Mohsen, H., El-Dahshan, E. S. A., El-Horbaty, E. S. M., & Salem, A. B. M. (2018). Classification using deep learning neural networks for brain tumors. *Future Computing and Informatics Journal*, *3*(1), 68-71.

[21]. Fatima, N. (2020). Enhancing Performance of a Deep Neural Network: A Comparative Analysis of Optimization Algorithms.

[22]. Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

[23]. Keras API reference-Optimizers: https://keras.io/api/optimizers/

[24]. Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, *12*(7).

[25]. Ney, H. (2017). Empirical investigation of optimization algorithms in neural machine translation. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 13-25.

[26]. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[27]. Cui, X., Zhang, W., Tüske, Z., & Picheny, M. (2018). Evolutionary stochastic gradient descent for optimization of deep neural networks. *arXiv preprint arXiv:1810.06773*.

[28]. Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.