Communication in Physical Science, 2021, 7(4): 388-397

A Deep Neural Network Approach for Cancer Types Classification Using Gene Selection

Florence Omada Ocheme, Hakeem Adewale Sulaimon and Adamu Abubakar Isah
Received: 30 November 2021/Accepted 24 December 2021/Published online:28 December 2021

Abstract: *Cancer classification research is one of the significant areas of exploration in the clinical field. Exact forecasting of various tumor types is an extraordinary challenge and giving an exact forecast will have incredible worth in giving better treatment to the patients. In recent years, many analysis-based investigations have led to the revelation of information on disease subtypes, that has generated high throughput innovations Lately, researchers have attempted to dissect a lot of microarray information for getting significant data that can be utilized in malignancy grouping. To accomplish this objective, one can utilize K-Nearest Neighbor, Neural Networks, Decision Tree, Support Vector a that would provide approaches needed to break down microarray information towards the choice of best separating quality called biomarker. These machine learning methodologies had the inherent ability to represent the time varying behavior of the underlying biological network that allows for a better representation of spatiotemporal input-output dependencies. Therefore, the exploitation of time series data regarding deep learning has to have become a valuable strategy for deciphering stochastic processes, such as gene expression and classification. Therefore, in this study, another intriguing strategy is introduced to improve the performance of neural networks utilizing deep autoencoder neural networks. This was accomplished through the choice of the first, relevant data, which was being extracted with a Deep Neural Network hidden layer used to train an autoencoder for the classification of the cancer malignancy based on the second stack autoencoder network. The outcome from the proposed experiment was evaluated with the current techniques. Overall, the proposed deep autoencoder accomplished classification accuracy of 99.2% as against the current*

Modified KNN and SVM which obtained 96.1% and 98.1% respectively.

Keywords: *Deep Learning, Artificial intelligence Neural Network, Autoencoder, K-Nearest Neighbor, Deep Recurrent Neural Network*

Florence Omada Ocheme
Department of Computer Science, Kaduna State University Kaduna State, Nigeria
Email: flor4cool@gmail.com
Orcid id: 0000-0001-6240-6629

Hakeem Adewale Sulaimon
Department of Computer Science, Federal College of Education Zaria, Kaduna State, Nigeria
Email: kasustudentwork@gmail.com
Orcid id: 0000-0002-1526-3829

Adamu Abubakar Isah
Department of Computer Science, Kaduna State University Kaduna State, Nigeria
Email: adamuaisah88@gmail.com
Orcid id: 0000-0001-6702-8735

1.0 Introduction

Cancer research is one of the major areas of interest in medical and pharmaceutical sciences because of the need to provide the solution for high life-threatening risk associated with it. However, precise prediction of different tumor types is a great challenge and the introduction of an accurate prediction model will add great value in the provision of better treatment of

cancer in patients. Over the past few decades, several experimental-based studies have been conducted to reveal the clue for the identification of cancer subtypes. Breakthroughs on high throughput technologies, such as gene expression profiling have invited opportunities for reform treatments by the analysis of the endogenous expression levels for thousands of genes in a single experiment. Generally, gene selection methods can assemble into four sets such as filter, wrapper, hybrid, and embedded methods (Shukla *et al*..2017). Fortunately, with the development of gene chip technology, the prospect of cancer classification and diagnosis at the gene expression level rises. However, the gene expression data typically have high dimensions and the samples of patients are few. Some of the genes may be immaterial to cancer classification. Therefore, to obtain excellent results in groupings of cancer treatment, we should select discriminatory genes and get a minor subset of genes (Luo *et al.*, 2009).

Several machine learning techniques have been largely explored for the classification of microarray datasets by several scientists, the use of new knowledge from the pathway level using deep learning has not been adequately addressed. These methods with the inherent capability to characterize the time varying behaviour of the underlying biological network allow for an improved representation of spatiotemporal input-output dependencies. Therefore, the exploitation of time series data regarding deep learning has been demonstrated to be a valuable approach for interpreting stochastic procedures, such as gene expression and classification (Halder and Kumar, 2019). In general, most of the reported methods still suffer from some deficiencies, such as high execution time and a stuck in local optima. Also, they do not attain reasonable classification results because they do not consider the minimization of the size of the selected target gene, and they need a maximum fitness assessment value and parameters for the adjustment. However, KNN, which was first presented by Fix and Hodges (1951), is one of the most influential and easy methods in microarray classification. However, it has several disadvantages such as the selection of K value, which influences the performance of KNN, v the necessities of distance calculation for those k neighbors and decreasing accuracy when multidimensional datasets are involved (Parvin *et al*., 2020). Therefore, an effort is also made in this investigation to address the existing gap due to low accuracy and high computational time associated with KNN (Tarek *et al*. (2017), The exploitation of time series data concerning deep learning has been demonstrated to be a valued strategy for deciphering stochastic processes, such as gene expression and classification. Several types of research using artificial intelligence methods have been uncovered to boost the classification accuracy as reported in the literature.

Vural and Subaşı (2015) employed data-mining methods to classify microarray gene expression data using gene selection by SVD and information gain, The investigational results showed that the proposed feature selection and dimension reduction gives improved classification performance in terms of the area under the receiver operating characteristic curve (AUC) and the forecast accuracy. In the work of Tarek *et al*. (2016), an effective cancer classification ensemble scheme is proposed. Ensemble classifiers rise not only the performance of the classification but also the confidence of the results. Muthuselvan and Sundaram (2016) evaluated five different algorithms using the data mining WEKA which includes Naïve Bayes, Zero R, One R, J48 and Random Tree algorithm for prediction of breast cancer using classification rule mining methods in blood test datasets. From the executed algorithm J48 algorithm is best, because the correctly classified instances are 86.3636% and also the MAE is 0.189 only. Ting and Sim (2017), proposed a self-regulated multilayer perceptron neural network for breast cancer classification (MLNN). Experimental results

prove that ML-NN can classify the input medical images as benign, malignant, or normal patients with accuracy, specificity, sensitivity and AUC of 90.59%, 90.67%, 90.53%, and 0.906 ± 0.0227 respectively. Also, Adrian *et al*. (2018) presented a classification system using Modified Backpropagation with Conjugate Gradient Polak-Ribiere and Ant Colony Optimization as the gene selection. it is found that the classification of MBP can achieve the F-Measure score of 0.7297. When joint with ACO as feature selection, the score rises by 0.8635. ACO demonstrated to optimize the classification technique of microarray cancer data excellently.

Rani and Ramyachitra (2018), proposed microarray cancer gene feature selection using a spider Monkey optimization algorithm and cancer classification using SVM. By conducting numerous experiments using ten different datasets, the results conclude that the proposed algorithm outperforms the other existing methods in classification accuracy and it chooses the fewer number of genes. Wu and Hicks (2021)] presented two different classifiers: Naive Bayes (NB) classifier and k-nearest neighbor (KNN) for breast cancer classification and evaluate their accuracy using cross-validation. Experimental results show that KNN gives the highest accuracy (97.51%) with the lowest error rate then NB classifier (96.19 %) but one of the major weaknesses is the high execution time that was observed by the KNN. Nawaz *et al*. (2018), proposed CNN to classify and diagnose breast cancer images from the public BreakHis dataset. The accuracy attained from the model is acceptable (73.68%) for the precise features of the input image. In the work of A new exciting method is presented. In their work. Ayyad *et al*. (2019) presented a method for the improvement of the functionality of deep neural networks using autoencoder neural networks. The proposed technique is based on variable Deep learning for gene expression data analysis. This approach This work also proposes a new classification method for gene expression data, which is called Modified k-nearest neighbor (MKNN) (Ayyad *et al*., 2019). Numerous experiments were conducted on six different gene expression datasets. Experimentations have revealed that MKNN in its both scenarios outperform (i) KNN, (ii) weighted KNN, (iii) support vector machine (SVM), (iv) fuzzy support vector machine, (v) brain emotional learning (BEL) in terms of classification accuracy, precision, and recall.

2.0 The Proposed Method

2.1 Data set preprocessing stage

The first part of the entire strategy is the dataset stage, which reads all the distinct datasets. Diverse operations are performed on those datasets and in the connections to the database including dataset loading and reading files. The proposed approach has been examined by six well-known datasets for ensuring the consistency of using our proposed approach to differentiate between different cancer types. A standard microarray dataset comprises a vast number of DNA particles spotted orderly on a solid material. Concerning the determining gene expression dataset stage, the preprocessing stage prepares the dataset for manipulation and it is considered an important step to handle gene expression data. Data normalization is used to remove systematic variation between samples, which is implemented so that each gene expression has a mean of zero and a variance of one. These procedures are essential to be performed before the classification process occurs.

2.2 Features selection and extraction stage using Deep Neural Network

In terms of high dimensional data classification, several features or genes do not include distinctive power but they decrease the classification performance. Consequently, there is a need for feature reduction to choose the greatest distinctive features or genes that discriminate various classes of cancers, Feature selection aims at the elimination of redundant and immaterial features for enhancing the classification procedure and decreasing the

computational cost. Feature selection is considered one of the most significant research areas in the machine learning field. The Deep neural network hidden layer first converts the data having diverse dimensions into a one-dimensional array or signal data to create a matrix of an array that fundamentally becomes an input to the model. This both allows to excerpt features from each period and to reduce the dimension of the input data.

2.3 Auto Encoder Classification Stage

Accordingly, the model performance can be directly influenced by the provided test dataset. To speed up the computational time of the system and rebuild the feature characteristics for raising the overall learning rate, the data is divided into different intervals (periods). Due to the features (size) of the data used in the experimental part, the trial-and-error method is employed to examine the best likely period for this training. Conferring to the proposed architecture, illustrated in Fig. 1, The gene features are extracted by the first hidden layer of the DNN, the output of the first deep neural network feature extraction level becomes an input to the second auto-encoder. This auto-encoder is trained in an unsupervised manner, producing a new data array with lower dimensions. The auto-encoder is trained by using the cost function as illustrated in equation 1. E value is regulated by employing the mean square error (MSE) approach.

$$E = \frac{1}{N}\sum_{N=1}^{n}\sum_{K=1}^{K}\left(X_{KN}\hat{X}_{KN}\right)^2 + \lambda * \Omega_{weights} + \beta\Omega_{sporsity} \quad (1)$$

Here, E is the loss rate (error rate), x is the input data, xÃ is the rebuilt data, l is the coefficient for the L2 Weight Regularization and b coefficient for the sparsity regularization, L is the number of hidden layers, n is the number of observations, k is the training data variables number. Besides, SoftMax classifier is trained to classify the output of the auto-encoder into labeled output. because it gives normalized class probabilities as outputs and is a preferred approach to maximize entropy between estimated and ground truth probabilities. In the last stage, the auto-encoders and SoftMax classifier are stacked and trained in a supervised manner as shown in the architecture illustrated in Fig. 1.

2.4 Implementation of the system

The developed deep autoencoder classifier model will be implemented using MATLAB R2018a. The computer to be used is a Toshiba laptop running on Windows 8 Operating System with 8GB RAM and Pentium ® Core i7 processor.

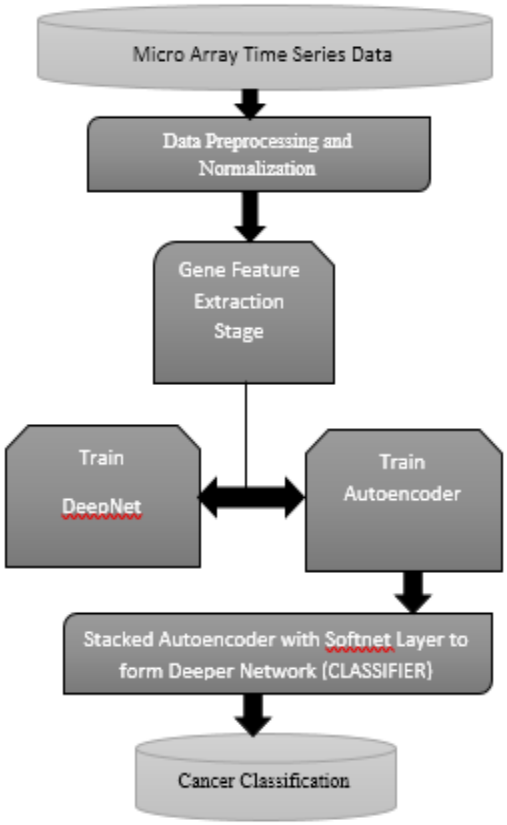

Fig. 1 Architecture of the proposed system

2.5 Dataset description

The evaluation of the suggested classification approach was carried out by the execution of the experimental results using six benchmark datasets for DNA microarray data. Table 1 shows the description of the dataset used in our research. All the datasets are related to a two-class classification problem and the datasets used in this study have several features

(thousands of genes), which are suitable for the display of the efficiency of the strategy in high dimensional data.

3.6 Choice of metrics

Testing and Evaluating a neural network are important stages of designing a classification or detection model. Hence, the proposed approach is evaluated using accuracy, precision, recall and F-Score to ascertain the best performing model against the state of the art.

Table 1: Data description

Dataset Name	Total Gene Per Instance	Classification type
Colon Turmor	2000	Turmor & Normal
Leukemia	7129	ALL & AML
Lung Cancer	125333	MPM & ADCA
DLBCL	4026	DLBCL & PL
Ovarian Cancer	1515	Turmor & Normal
Prostate Cancer	12600	Turmor & Normal

4.0 Results and Discussion

4.1 Implementation of the System

The developed deep learning classifier model was implemented using MATLAB R2018a. The computer used is HP laptop running on Windows 10 Operating System with 12GB RAM, 750 GB HDD and Pentium ® Core i7 processor.

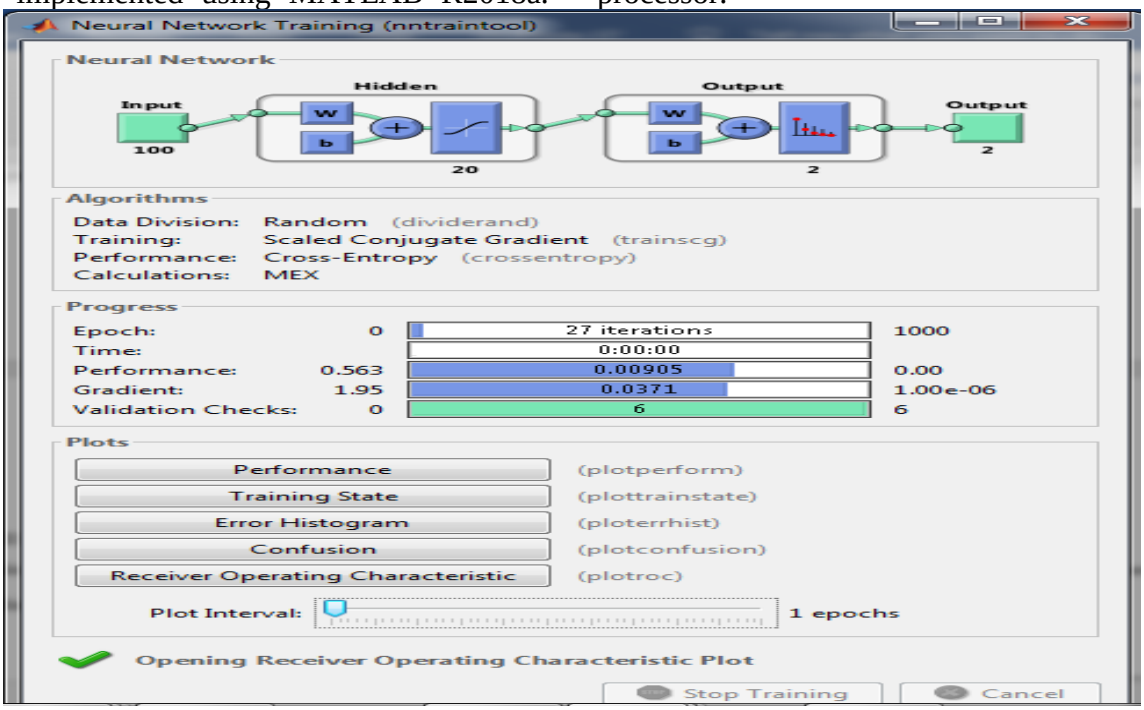

Fig 2: Training performance toolbox for the proposed algorithm

This study addresses a new framework based on cancer classification using a microarray time series dataset and deep sparse auto-encoder architecture. The proposed techniques were tested on the benchmark datasets against two widely classified approaches, namely, SVM and KNN. For each one, the precision, recall, accuracy, and computation time were recorded.

The experiments were performed on a machine with Intel corei7. the simulation result was analyzed and compared with three different phases including accuracy, training performance and computational time. The figure below was obtained from the MATLAB simulation environment portraying deep learning architecture such as the parameters configuration, training progress, number of iterations and epochs and validation accuracy for the proposed deep learning on ovarian dataset a case study. Also, Fig, 3 shows the validation MSE for the proposed model on the ovarian dataset. The training algorithm used in the work to train the proposed model is the Bayesian regularization backpropagation algorithm which converged after 21 epochs, and it showed stability (no increase after converging) and no overshoot (no increase before converging), as shown in Fig.3.

Fig. 4 shows the ROC curve of the proposed algorithms on the ovarian dataset it describes the ROC (Receiver Operating Characteristics), A ROC curve was constructed by plotting the true positive rate (TPR) against the false positive rate (FPR). The true positive rate is the proportion of observations that were correctly predicted to be positive out of all positive observations (TP/ (TP+ FN)). Similarly, the out of all negative observations (FP/ (TN + FP)) false-positive rate is the proportion of observations that are incorrectly predicted to be positive. Our ROC in the figure above shows that the classifier classified both positive and negative samples are correctly classified, hence a perfect classifier.

Also, Fig. 5 shows the confusion matric of the proposed algorithms on the ovarian cancer dataset., the green diagonal cells represent the number and percentage of correct classifications by the trained network while the red diagonal cells stand for the incorrect classifications, which are 0 for each target class and each output class. The lower right ash square illustrates the overall accuracy. Therefore, the proposed model achieved over 96.8% on average which indicates a very good classification accuracy.

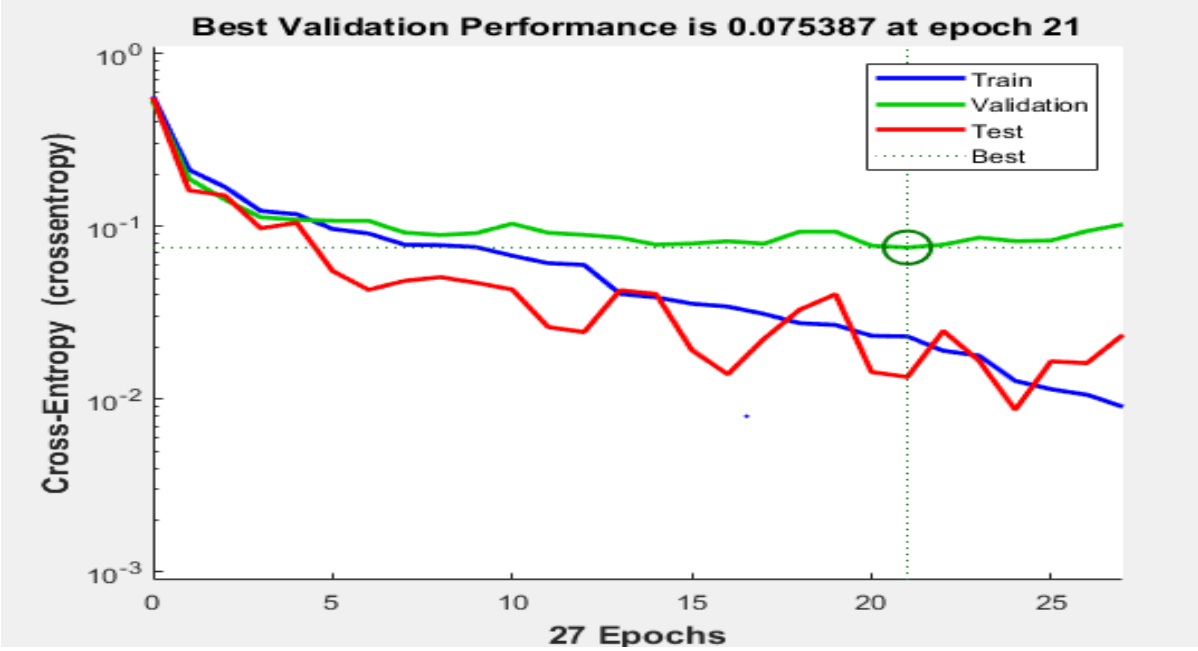

Fig. 3: The validation MSE for the proposed model on an ovarian dataset

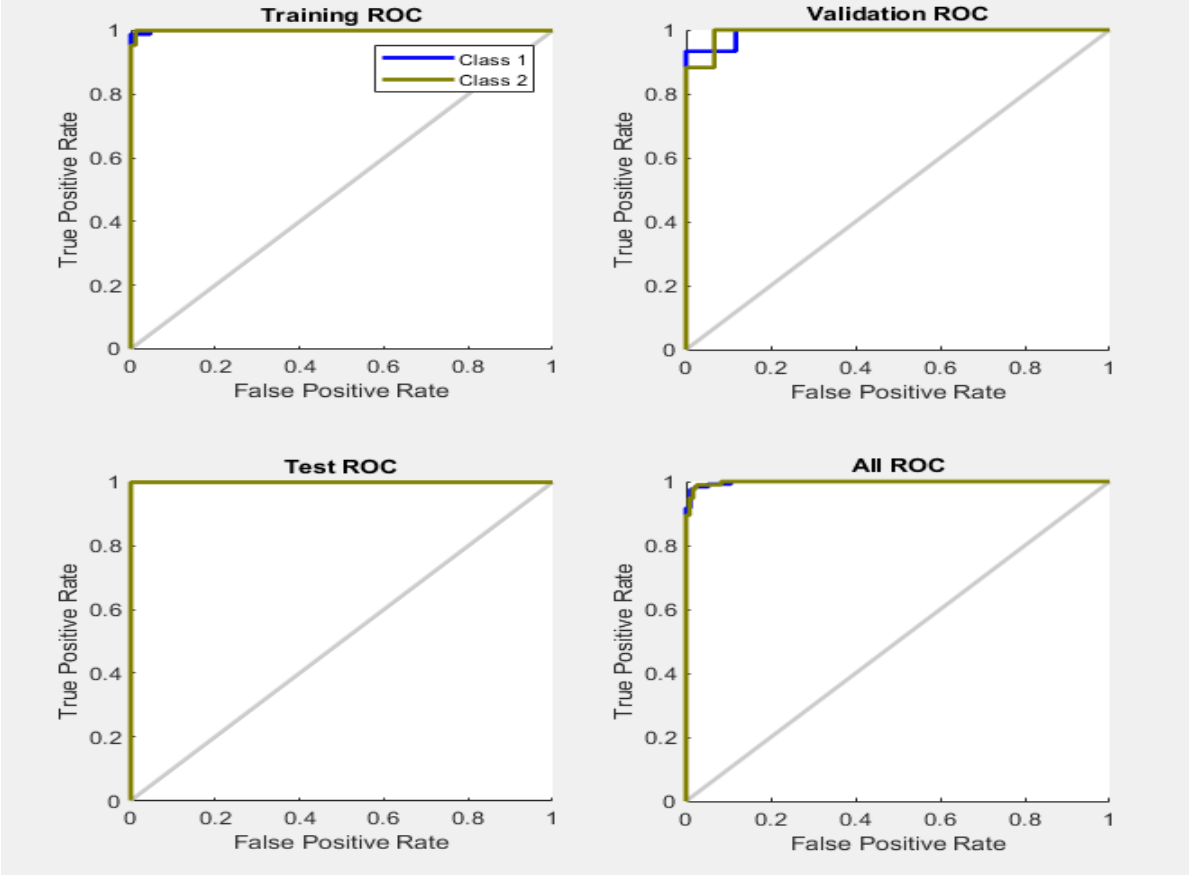

Fig 4. shows the ROC curve of the proposed algorithms on the ovarian dataset

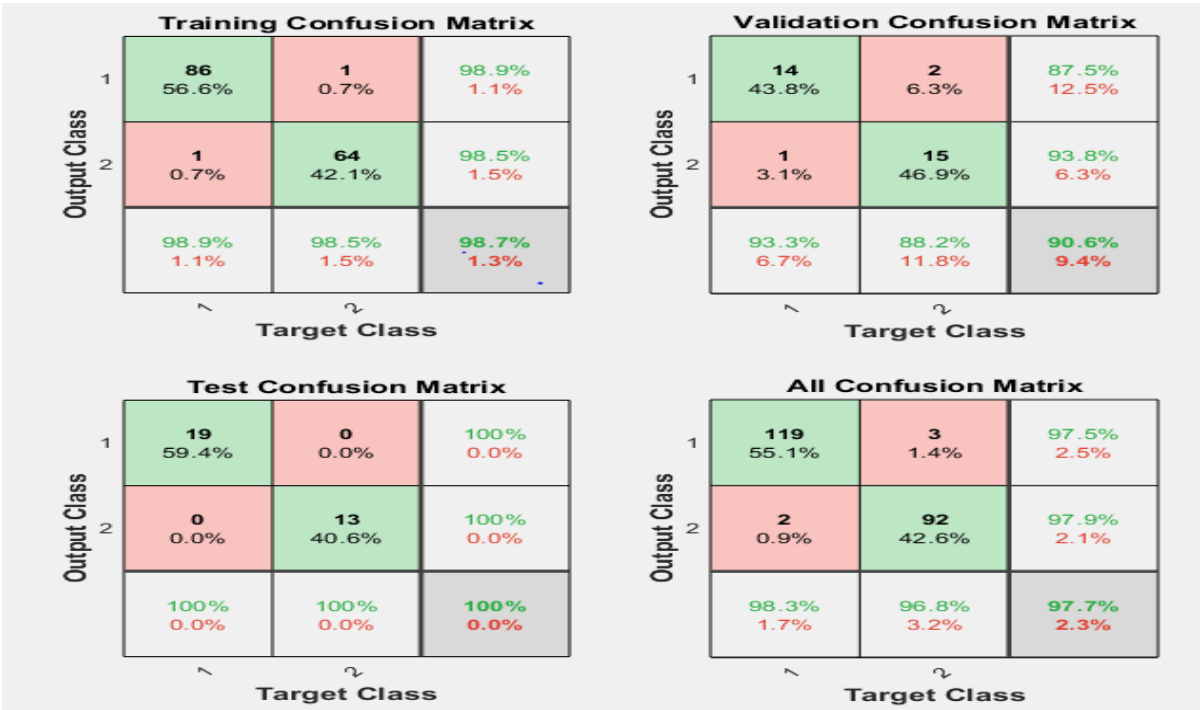

Fig. 5: Confusion matrix

4.3.1 Classification accuracy

Table 1 shows the classification accuracy of the simulation results obtained after performing iterations on each of the six-benchmark cancer datasets. However, Fig. 6 shows the classification accuracy across the six data sets It can be deduced from Table 1 that the proposed model achieved a higher classification accuracy when compared to the existing SVM and modified KNN. For classification accuracy, the higher the value in decimal place, the better classification is obtained. The proposed deep network achieved the best accuracy of 99.8%, 99.1% and 99.2% for colon tumor, ovarian cancer and prostate cancer respectively as against the existing modified KNN and SVM which obtained the best result of 96.1% and 98.3% for lung cancer and ovarian cancer respectively. In general, the proposed model achieved a better classification accuracy as against the conventional techniques.

4.3.2 Precision and recall

Table 2 below shows the values obtained for both precision and recall from the simulation results obtained after 100 iterations.

Table 1: Classification accuracy across the six datasets

Cancer Type	Proposed Model	MKNN	SVM
Colon Tumor	99.8	96.1	98.1
Leukemia	88.4	96.2	98.3
Lung Cancer	95.2	96.1	98.1
DLBCL	96.1	96.2	98.2
Ovarian Cancer	99.1	96.2	98.3
Prostate Cancer	99.2	96.1	98.1

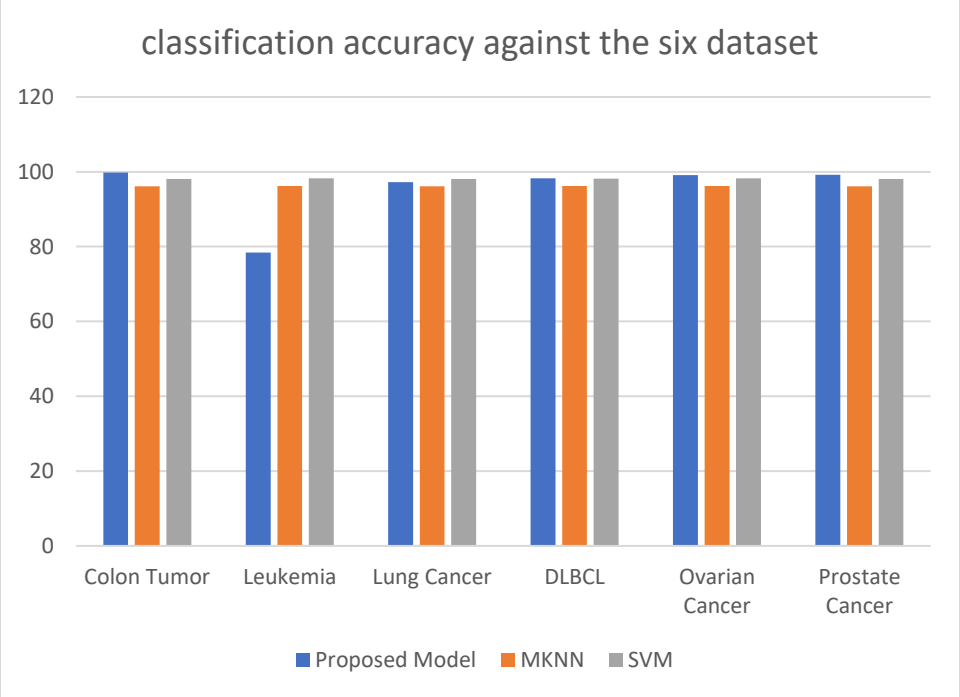

Fig. 6: Graph showing the Classification Accuracy against the six dataset

Table 2: Classification precision and recall across the six datasets

Cancer Type	Proposed Model		MKNN		SVM	
	Precision	Recall	Precision	Recall	Precision	Recall
Colon Tumor	0.788	0.866	0.601	0.952	0.880	0.573
Leukemia	0.998	0.788	0.761	0.804	0.782	0.690
Lung Cancer	0.875	0.976	0.922	0.733	0.878	0.783
DLBCL	0.992	0.973	0.684	0.862	0.781	0.876
Ovarian Cancer	0.851	0.899	0.876	0.891	0.683	0.893
Prostate Cancer	0.992	0.912	0.523	0.681	0.581	0.877

From the results presented in Table 2, it is noticeable that the proposed deep learning attains a first place in most cases and MKNN attains second place and is only inferior to SVM on the benchmark datasets. Furthermore, it is also clear that the proposed deep learning attains the highest level of performance by obtaining a value very close to 1 in recall and precision.

4.3.4 Average computational time in seconds

Finally, a comparison in computational time is displayed in Table 3, which reveals that MKNN and SVM perform too slow.

As expected, KNN is a lazy learner. So, more time is consumed throughout the testing stage as nearly all calculations are executed during the testing stage. On the contrary, SVM consumes the longest time as it executes classification tasks by building the best hyperplane in a multidimensional space by increasing the margin as possible.

Table 3:Average computational time in seconds on tumor dataset

Iterations	Proposed Model CPU TIME(S)	MKNN CPU TIME(S)	SVM CPU TIME(S)
100	43	55	67
500	185	243	326

Table 4 shows that the computation time of MKNN and SVM is close, but the proposed deep learning attains the lowest time. The above experimental results imply that MKNN and SVM are effective for classifying high dimensional data even without dimensionality reduction. It can effectively obtain a high degree of classification accuracy. The overall performance of the proposed deep learning is superior to the MKNN and SVM. Additionally, it is worth noticing that MKNN and SVM also perform comparatively good across all six datasets whether in classification metrics or computational time.

5.0 Conclusion

In this research, a new interesting technique is presented to enhance the functionality of deep neural networks using autoencoder neural networks. The proposed technique is based on variable Deep learning for gene expression data analysis. The main idea is to classify the samples based on deep learning. For classification accuracy. On average, the proposed deep autoencoder achieved an accuracy of 99.2% as against the existing Modified KNN and SVM which obtained 96.1% and 98.1% respectively.

Despite the promising results attained via deep architectures, there remain numerous unresolved challenges facing the medical application of deep learning to health care. In our opinion, we recommend that research directions should comprise both algorithms to explain the deep models (i.e., what drives the hidden units of the networks to turn on/off along the process as well as approaches to support the networks with current tools that explain the forecasts of data-driven systems.

5.0 References

Shukla, A. K., Singh, P. & Vardhan, M. (2020). Gene selection for cancer types

classification using novel hybrid metaheuristics approach.. *Swarm and Evolutionary Computation*, 54, 100661,, doi:10.1016/j.swevo.2020.100661.

Tarek, S., Elwahab, R. A. & Shoman, M. (2017). Gene expression based cancer classification. *Egyptian Informatics Journal*, 2017. 18(3): p. 151-159.

Halder, A. & Kumar, A. (2019). Active learning using rough fuzzy classifier for cancer prediction from microarray gene expression data. *Journal of Biomedical Informatics*, 2019. 92, 103136, https://doi.org/10.1016/j.jbi.2019.103136

Fix, E. & Hodhes, J. L. (1951). *Discriminatory analysis, nonparametric discrimination: consistency properties*. Report Number, 4, Project Number 21-49004, UDAF School of Aviation Medicine, Radolph Field, Texas.

Luo, W., Wang, L., & Sun, J. (2009). *Feature selection for cancer classification based on support vector machine.* Paper presented at the 2009 WRI Global Congress on Intelligent Systems.

Piao, Y., M. Piao, and K.H. Ryu, Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles. Computers in biology and medicine, 2017. 80: p. 39-44.

Rani, R.R. & Ramyachitra, D. (2018). Microarray cancer gene feature selection using spider monkey optimization algorithm and cancer classification using SVM. *Procedia computer science*, 143, pp. 108-116.

Ayyad, S.M., Saleh, A. I. & Labib, L.(2019). Gene expression cancer classification using modified K-Nearest Neighbors technique. *BioSystems*, 176, pp. 41-51.

Parvin, H., Alizadeh, H. &. Minati, B. (2010). A modification on k-nearest neighbor classifier. Global Journal of Computer Science and Technology, 2010.

Vural, H. & Subaşı, A. (2015). Data-mining techniques to classify microarray gene expression data using gene selection by svd and information gain. *Modeling of Artificial Intelligence*, 2, pp. 171-182.

Tarek, S., Elwahab, R. A. & Shoman, M. (2016). *Cancer classification ensemble system based on gene expression profiles*. in 2016 5th International Conference on Electronic Devices Systems and Applications (ICEDSA). 2016. IEEE.

Muthuselvan, S. & Sundaram, S. (2016). *Prediction of breast cancer using classification rule mining techniques in blood test datasets*. in 2016 International Conference on Information Communication and Embedded Systems (ICICES). 2016. IEEE.

Ting, F. & Sim, K. (2017). Self-regulated multilayer perceptron neural network for breast cancer classification. in 2017 International Conference on Robotics, Automation and Sciences (ICORAS). 2017. IEEE.

Adrian, D. & Annisa, A. (2018). Cancer detection based on microarray data classification with ant colony optimization and modified backpropagation conjugate gradient polak-ribiére. in 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA). 2018. IEEE.

Nawaz, M., Sewissy, A. A. & Soliman, T. H. A. (2018). Multi-class breast cancer classification using deep learning convocational neutral network. *Journal of Advanced Computer Science and Application, 9, 6, doi.* 10.14569/IJACSA.2018.090645 .

Wu, J. & Hicks, C. (2021). Breast Cancer Type Classification Using Machine Learning. *Journal of Perspective Medicine*, 11, 2, 61, doi: 10.3390/jpm11020061.

Conflict of Interest

The authors declared no conflict of interest