A Comparative Study on Classification Methods for Renal Cell and Lung Cancers Using RNA-Seq Data

BULENT HAZNEDAR and NIHAT Y. SIMSEK

1Computer Engineering Department, Gaziantep University, 27310 Gaziantep, Turkey
2Institute of Natural Science, Hacettepe University, 06800 Ankara, Turkey

Corresponding author: Bulent Haznedar (haznedar@gaziantep.edu.tr)

ABSTRACT Nowadays, The gene expression analysis gains a significant research interest and plays an important role for the classification and diagnosis of cancer types. In such research studies, the main difficulty is the processing time consumed due to numerous numbers of genes to be classified in human cell. RNA-Seq is a novel technology which enables researchers to obtain reliable knowledge in the analysis of numerous number of genes, so that can be effectively used for cancer classification. In this paper, commonly-used deep learning model based on deep neural network architecture has been proposed and utilized to analyze lung and renal cell cancer RNA-Seq datasets taken from The Cancer Genome Atlas (TCGA). The proposed method is compared with commonly-used other classical machine learning algorithms including decision trees (DT), random forests (RF), support vector machines (SVM) and artificial neural network (ANN) in terms of performance and accuracy for the same datasets. This study also presents the effects of different optimizers to the performance of deep learning algorithms. As a result, the proposed deep learning model have yielded the highest accuracy of 96.15% on renal cell and 95.54% on lung cancer data. It is found that the proposed deep learning model is very successful in classification of RNA-Seq datasets with large number of features compared. When results are compared with a previous study in literature which also analyses the same datasets, the proposed deep learning model outperforms the all other methods in various metrics.

INDEX TERMS Classification, deep learning, gene expression, machine learning, RNA-Seq.

I. INTRODUCTION

Cancer is primarily a genetic disease and generally starts with a series of mutations on a single cell that becomes an abnormal cell. Then the abnormal cell divides uncontrollably and can spread throughout the tissues, organs, or body. Gene mutations associated with cancer can be inherited from parents or occurred through somatic mutations. Diagnosis and classification of cancer by gene expression has a significant importance at this point. Gene expression is the process that contains the necessary information for the formation of a gene. Gene expression shows the activation status of a gene during protein production.

Microarray gene expression is one of the most well-known tool used in laboratory for detecting the expression of many genes simultaneously. The data collected from microarrays can be effectively used for diagnosis and classification of human cancer [1]. However, there exist many limitations related to usage of the microarray technology. Specifically, some of these are the followings: Firstly data variability especially for genes with low expression level, secondly unequal labelling efficiency of fluorescent dyes, thirdly small sample amounts which might limit replication, and lastly lack of information related to the protein expression levels and function. In addition, usage of microarray technology might limit the researchers for detecting transcripts that correspond to existing genomic sequencing information [2]. Because of these limitations, more recent technology became more popular for analysing the gene expressions is RNA-Seq and it...
has been found that the RNA-Seq technology has a few main advantages over the microarray technology, so that RNA-Seq technology has started to become the major principle and commonly-used method in gene-expression research studies [3]. RNA-seq experiments performs well in investigation of both known and unknown transcripts. Most of the time, so that RNA-seq can be a good candidate for discovery-based studies. RNA-seq experiments can be updated once new sequence information obtained while microarrays are limited to the reference information exist during the production. Because, DNA sequences directly overlap with the unique regions of the genome, RNA-seq has the capability to remove the noises easily from the experiment. Lastly, RNA-seq is capable of quantifying large dynamic ranges of expression levels by using absolute values rather than relative ones. In RNA-Seq technology, Next-Generation Sequencing (NGS) is used for determining the RNA amount and sections by analyzing the gene expression transcriptome stored in human RNA [4]. RNA-Seq can help researchers to determine which genes are enabled, their expression level, and they are activated or deactivated under what circumstances in the cell [5]. This enables researchers to gain a deeper understanding of cell genetics and evaluate changes that can be used for diagnosis of a disease.

Many studies have been conducted on the diagnosis and classification of cancer diseases by analyzing gene expression data. These studies usually comprise SVM and other machine learning algorithms. Recently, deep learning algorithms have also been widely used for the purpose of analyzing gene expression data. In this study, we have developed a deep learning model which can be considered as the recent developing technology to analyze lung and renal cell cancer RNA-Seq datasets.

Deep Learning is a branch of machine learning and uses computational models which are formed of multiple processing layers to learn representations of data at high-level of abstraction. Very complex functions can be easily learned with sufficient combination of transformations. For classification tasks, higher representation layers strengthen the aspects of input that is very crucial to discriminate and suppress irrelevant variations. Another potential feature of deep learning is changing manual features by using effective algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction [6].

A semi-supervised deep learning model used by Xiao et al. [7] for the cancer prediction by using RNA-Seq dataset. They developed a stacked sparse auto-encoder (SSAE) based method then SSAE was tested on three different cancer RNA-Seq datasets. It was concluded that the developed method yielded better classification performance in various metrics.

In the present study, a deep learning model based on deep neural network (DNN) architecture is developed for the classification of RNA-Seq RCC and lung cancer datasets. The challenge in the present study concerns the complexity and high dimensionality of RNA-seq data. In order to obtain a high level of accuracy in classification, a feature selection process specifically wrapper method is applied to the high-dimensional RNA-seq data to reduce its dimension by selecting only the best features and removing the irrelevant features. Afterwards, DNN method with capability of multiple processing layers to classify RNA-seq data at high-level of abstraction. In this scope, the effects of different optimizers are investigated and their performances are analyzed in detail. Also, efficiencies of the different models are checked against four different classical machine learning algorithms including DT, SVM, and RF and Artificial Neural Network (ANN), which are commonly-used in literature. A feature selection method specifically wrapper method is applied on these datasets before utilizing the machine learning and DNN models mentioned above. From the obtained results, It is found that the proposed DNN-Adadelta model is the most successful among the all other methods.

Briefly, the effective RNA-seq technique is adopted for the proposal of a deep learning model based on deep neural network architecture. The effects of different optimizers such as SGD, RMSProp, Adagrad, Adadelta, Adam, Adamax and Nadam are investigated to develop an optimum DNN model. AdaDelta optimizer is found to be the most successful in classifying RNA-Seq cancer datasets analyzed in this study. It is demonstrated that the developed deep learning model can be successfully used for the analysis of RNA-Seq data for specific cancer types.

II. LITERATURE REVIEW

In literature, there exist several studies which have analyzed gene expression datasets by using microarray technology. Huang [8] applied three different classification algorithms to four different cancer microarray datasets. DT, SVM and k-nearest neighbors (KNN) were used as the classification algorithm for the analysis of hepatatox, colon cancer, lymph cancer and leukemia microarray datasets. At the end of that study, DT method on leukemia data yielded the highest accuracy rate of 96.6%.

In another study, leukemia, brain tumor, prostate and colon cancer datasets were analyzed with the used of four different classification algorithms [9]. Firstly, gene selections were made in datasets before applying the classification algorithms. Datasets were divided into subsets of five, ten, twenty, fifty and one hundred genes. Lastly, classification algorithms were applied to these subsets and it was found that Naive Bayes algorithm achieved 91.1% best accuracy rate on colon cancer among all methods used.

Tran et al. [10] have presented a computational method for prediction of the tumor tissues by using high-throughput miRNA expression profiles and found that the informative miRNAs show strong distinction of expression level in tumor tissues. Using the microarray dataset used by Gloub et al. [11], they classified samples as tumors and normal cells. This dataset has 223 samples with 151 mRNA properties. In that study, SVM with 3 different kernel types including Linear, Polynomial, and Radial Basis Function
(RBF) were used. As a result of performed classification with RBF, Linear, and Polynomial kernel types have revealed an accuracy rate of 92.00%, 95.00%, and 93.00%, respectively.

Zararsiz et al. [12] applied seventeen different classifier algorithms to four different RNA-Seq datasets including the cervical, Alzheimer’s, renal cell cancer (RCC) and lung cancer RNA-Seq datasets; SVM and Random Forest (RF) yielded the best accuracy rates. At the end of that study, SVM is found to be most successfully with an accuracy rate of 93.5% for RCC and 94.8% for lung cancer. In the present study, we have analyzed the same RCC and lung cancer datasets, so we have directly compared our results to their results (see Sec. IV-C).

Tan and Cahan [13] presented their own model named SingleCellNet for the classification of single-cell RNA-seq data. This tool compared favorably to other methods in terms of sensitivity and specificity.

Alquicira-Hernandez et al. [14] developed a new method named scPred method which was used for classification of RNA-Seq dataset by using combination of unbiased feature selection and machine-learning based on prediction method. Simsek and Haznedar [15], [16] applied DNN to classify the same RCC dataset. In the previous work, DNN was optimized by using only RMSProp algorithm. The results was compared to the classical RF method. A satisfactory results were obtained for classification of RCC dataset.

III. MATERIALS AND METHODS

In this section, analyzed datasets are introduced then feature selection methods, classification algorithms, and evaluation criteria for the results are explained in detail.

A. DATASET DESCRIPTION

Two different datasets for Renal cell and lung cancer are analyzed in this study. Detailed descriptions and analyses of these datasets are given in the following subsections.

Renal Cell Cancer Dataset: There are many datasets available for researchers to download and analyze in The Cancer Genome Atlas (TCGA) which is a comprehensive community resource platform. Renal cell cancer (RCC) dataset is obtained from TCGA in the form of RNA-Seq dataset [17]. This dataset consists of 1,020 RCC samples with 20,531 RNA transcripts for each. This RNA-Seq data has 606, 323, and 91 specimens from the kidney renal papillary cell (KIRP), kidney renal clear cell (KIRC) and kidney chromophobe carcinomas (KICH), respectively. These three types of cancers are the most well-known subtypes of RCC (account for nearly 90%-95% of the total malignant kidney tumors in adults) and treated as three different classes in that study [18].

Lung Cancer Dataset: Another RNA-Seq dataset obtained from TCGA is lung cancer dataset [17]. There are 1,128 samples in this dataset and each sample includes 20,531 transcripts. There are two different classes that are lung adenocarcinoma (LUAD) and lung squamous cell with carcinoma (LUSC) with 576 and 552 class sizes, respectively. In this paper, these two lung cancer types are treated as two different classes. Detailed description of RCC and Lung Cancer Datasets are presented in Table 1.

B. FEATURE SELECTION AND WRAPPER METHOD

When applying machine learning algorithms to a problem, it is crucial to create a suitable model for the problem. Also, one of the most important factors affecting the model’s prediction performance is the number of input features used in the model. Feature selection is the process of decreasing the number of input features by identifying and getting rid of features with negligible effects on the results. Feature selection is very important process because the high number of input variables increases time and cost of calculation and decreases the model’s training performance [19].

In this study, in order to increase the performance of the developed model, feature selection is applied on the datasets, specifically wrapper method is chosen and used to reduce the number of genes count.

| Dataset     | Genes Count | Number of Samples | Subclass | Total |
|-------------|-------------|-------------------|----------|-------|
| Renal Cell  | 20,531      |                   | KIRP     | 606   |
| Cancer      |             |                   | KIRC     | 323   |
|             |             |                   | KICH     | 91    |
| Lung Cancer | 20,531      |                   | LUAD     | 576   |
|             |             |                   | LUSC     | 552   |

FIGURE 1. Wrapper method algorithm.
trained classifier A is tested on each subset. Then results are compared using by cross-validation method.

Wrapper method consumes more computational time than other feature selection techniques but it is better to have a good bias which is more suitable for learning algorithm to perform a better performance.

C. CLASSIFICATION AND DEEP LEARNING

Classification is the process of separating the elements of a given dataset according to its category. The classification process can be applied on both structured and unstructured datasets. The purpose of the classification models is to find out an input data belongs to which class and the main purpose of using a model is to be able to classify a new data introduced later to analyzed dataset. In this paper, it is aimed to classify the data by using proposed deep learning algorithm.

Deep learning can be utilized to make a great progress even for the problems that cannot be solved for years by using classical machine learning algorithms. Deep learning has been considered to be a master in solving very complex problems including high-dimensional data for various scientific, business and government related fields. Additionally, various studies show that the deep learning models usually yields better results and performances than machine learning algorithms in research studies including image and speech recognition [20], [21], drug molecules [22], analysis of particle accelerator data [23], reconstruction of brain circuits [24], and prediction of mutation effects in non-coding DNA on gene expression and disease [25].

There are different types of deep learning architectures like deep neural networks, deep belief networks, recurrent neural networks, and convolutional neural networks. Comparison of architectures of ANNs and Deep Learning is depicted in Fig. 2. In this study, a deep neural network (DNN) model with seven different optimizers is proposed and their performances are evaluated.

DNN is in fact an ANN with several hidden layers of units across the input and output layers [26]. DNN can be used to treat complex non-linear relationships as ANN, but DNN have the extra layers which allows feature combinations from lower layers. Hence, DNN have more capability to create models for complex data with less units than networks designed similarly [27]. DNNs are generally aimed to function as feed-forward networks and they can be discriminatively trained with the standard back-propagation algorithm which is used to update weights (W) in accordance with Eq. 1.

$$W(k) = W(k-1) - \eta_k \frac{\partial E(k)}{\partial W(k-1)}$$

where, $$\eta_k$$ denotes the learning rate and E is the error function for the $$k^{th}$$ training iteration. The selection of the cost function depends on parameters like the learning model (supervised, unsupervised etc.) and the activation function. For instance, given that supervised learning is applied on a multiclass classification problem, softmax function can be chosen as the activation function and cross entropy function can be used as cost function. Mathematically, the softmax function can be expressed with Eq. 2.

$$P_j = \frac{\exp(x_j)}{\sum_k \exp(x_k)}$$

where, $$P_j$$ represents the probability of class (output of the unit j) and $$x_j$$ and $$x_k$$ represent the total input to units j and k of the same level, respectively. Cross entropy (cost function in a supervised learning on multiclass classification problems) is obtained from Eq. 3.

$$c_r = \sum_j d_j \sqrt{\log(P_j)}$$

where, $$d_j$$ represents the target probability for output unit j and $$P_j$$ is the probability output for j after applying the activation function [28]. DNN-based regression is a good classifier which is able to learn features grabbing geometric information too. DNN eliminates the limitations in creating a model in terms of obtained parts and their relations and this contributes to learn a wide range of objects. The model comprises of multiple layers and each has a rectified linear unit for a non-linear transformation. Some of the layers are convolutional, whereas others are fully connected and these convolutional layers have an extra max pooling. The network is trained to reduce L2 error for the prediction of mask ranging over the whole training set including bounding boxes represented as masks [28]. In this paper, a model developed using the mentioned advantages of deep learning are compared with other classical machine learning algorithms.

AdaDelta is developed as an extension of the Adagrad algorithm. In contrast to Adagrad algorithm, the continually decreasing learning rate and manual selection of a global learning rate issues are avoided in AdaDelta. In AdaDelta optimizer, the denominator is formed of the squared gradients from the beginning for each iteration, so that terms are positive which primarily helps to avoid cancellation effects in the cumulative sum. As the sum gets accumulated, it gets increasing and after a certain iteration, change in the learning rate becomes infinitesimally small and the algorithm gets converged. Because of this reasons, AdaDelta is utilized for the optimization of the DNN network in this study.
TABLE 2. A confusion matrix for a binary classification problem.

| Actual Values | Predicted Values |
|---------------|------------------|
|   Positive    |    Negative      |
|   TP          |    FN             |
|   TN          |                  |

D. EVALUATION STRATEGIES

In this section, evaluation strategies used to quantitatively analyze the performance and accuracies of the results from different models are explained in detail.

Mean Absolute Error (MAE): The MAE finds the average magnitude of errors in a series of estimates, regardless of their direction. It can calculate the accuracy for continuous variables as well. The MAE is the average of the absolute values of the differences between the estimate and the coincident observation relative to the true value. MAE is a linear score; this means that all individual differences are on average equal weights [29]. The mean absolute error can be calculated by using Eq. 4.

\[
MAE = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n}
\]

where \( |x_i - y_i| \) corresponds to the absolute error of \( i \)th prediction \( x_i \) with respect to the \( y_i \) target value.

Root Mean Squared Error (RMSE): RMSE is a quadratic scoring principle which can be also used to calculate the error’s average magnitude. It is the square root of the mean differences in squares between prediction and real observation. The root mean square error is expressed with Eq. 5.

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}}
\]

where, \( x_i \) represents the prediction class and \( y_i \) represents the result of truly classified values. RMSE is a measure that can be used to compare the errors in predictions of different models in terms of accuracy [30].

Predicted results of a classification problem’s summary is called confusion matrix. The key to the confusion matrix is to summarize and classify the number of true and false estimates by the degree of counting. From this point, the confusion matrix tells the ways in which the classification model is “confused” while making predictions. It provide also the information not only into the failures made by a classifier, but more importantly the type of the resulting failures. Stucture of a confusion matrix for a binary classification problem is presented in in Table 2.

The terms which are given in Table 2 are described as below,

- True Positive (TP): Actual class label is positive, and predicted label is positive.
- False Negative (FN): Actual class label is positive, but predicted label is negative.
- True Negative (TN): Actual class label is negative, and predicted label is negative.
- False Positive (FP): Actual class label is negative, but predicted label is positive.

One may calculate the classification accuracy via Eq. 6 by using values presented in the confusion matrix.

\[
Accuracy(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100
\]

IV. RESULTS AND DISCUSSIONS

A. FEATURE SELECTION RESULTS

For the purpose of selecting the most-relevant genes in datasets, wrapper method has been utilized. Python programming language has been used within TensorFlow environment to apply a correlation method for gene selection properly. Random Forest Regressor utilized as the estimator in the training model with statistical metric R-square value. The diagram of the developed feature selection method is depicted in Fig. 3.

Dataset have been divided into two sets of data to be used for training and test purposes. Classification of RNA-seq data with small number of samples similar to the datasets in this study is a key issue in bioinformatics and statistics.
B. Haznedar, N. Y. Simsek: Comparative Study on Classification Methods for Renal Cell and Lung Cancers Using RNA-Seq Data

Usage of validation set requires large number of samples to come up with high accuracy estimates for data splitting. So that the usage of validation set is practical and rewarding only for the datasets with large number of samples. Because of these reasons, a random sampling is adopted as primary methodology for data splitting in this study. Quantitatively, 80% of dataset has been used for training and 20% for testing samples. K-fold cross validation value has been taken as five to increase the dataset variance. After feature selection has been performed, the best suited 50 genes have been determined and to be used as input to the proposed classification model. These 50 genes are listed in Table 3 and Table 4 for the RCC and Lung Cancer datasets.

Upon selection of these genes, five different classification methods applied on preprocessed samples with new subsets of original RNA-Seq datasets.

**B. CLASSIFICATION RESULTS**
Following up the gene selection, the classification algorithms have been applied on RCC and lung cancer datasets. Firstly, classical machine learning algorithms have been applied for comparison purpose and later the deep learning model has been studied. Investigated classical machine learning algorithms are DT, RF, and also three different types of SVM and ANN. Many attempts are applied to generate the optimum DNN model. As a consequence of attempts, the developed DNN model yields the optimum results with 7 layers where the softmax function is applied to the output layer. Afterwards, RMSProp, SGD, Adagrad, Adadelta, Adam, Adamax and Nadam optimizers are applied for the optimization of the DNN model. The control parameters of the developed DNN model are as the following: the epoch number is 200, the loss function is crossentropy, the weight decay is 0.00005 and the dropout rate is 0.5 to avoid the network from overfitting. Later, these classification algorithms have been applied to the lung cancer dataset by reserving 70% for training and 30% for the test of the dataset. The LUAD cancer type is designated by 0, while the LUSC cancer type is designated by 1. After using these classification models on the train set and MAE value and root mean RMSE have been calculated for evaluation purposes. Comparisons of the results of the classical algorithms, DNN models with different optimizers, and classical algorithm vs. the best DNN model for the lung cancer dataset are tabulated in Tables 5-7, respectively. Loss and accuracy change graph of the proposed deep learning model for lung cancer RNA-Seq dataset during the training are shown in Fig. 4 and Fig. 5, respectively. Also the developed model and classical machine learning algorithms have been applied to the renal cell cancer dataset by reserving 70% for training and 30% for test of the dataset. This percentage values are determined based on many attempts to improve the efficiency of the classification algorithms.
method. The KICH, KIRC and the KIRP cancer types are designated by 0, 1 and 2, respectively. All models have been applied on the train set; Accuracy, MAE, and RMSE have been calculated for the classification test results from each model. Comparisons of the results of the classical algorithms, DNN models with different optimizers, and classical algorithm vs. the best DNN model for the lung cancer dataset are tabulated in Tables 8-10, respectively.

Loss and accuracy change graph of the developed deep learning model for RCC RNA-Seq dataset during the training phase are shown in Fig. 6 and Fig. 7, respectively.

The confusion matrix provides the information not only into the failures made by a classifier, but more importantly the type of the resulting failures. In this context, the performance of proposed DNN-Adadelta method is also evaluated by using confusion matrix. The confusion matrices for the lung cancer and RCC dataset are presented in Table 11 and Table 12, respectively.

As depicted from Table 11 and Table 12, smaller scale off-diagonal FP and FN values are obtained relative to the diagonal TP and TN values as desired.

C. DISCUSSION

Based on the test accuracy, RMSE and MAE values in Table 5 and Table 8, it clearly seen that RF yields the best accuracy of 93.51% and 91.83% values for both lung cancer and RCC among the classical machine learning algorithms, respectively. The comparison of MAE and RMSE values
also indicates that RF gives the lowest error values for both datasets. Hence, it can be concluded that RF is the best classification algorithm among the investigated machine learning algorithms for RNA-Seq datasets in this study. Three different kernel types of SVM have been applied to datasets and their results are tabulated in Table 5, the developed SVM with RBF kernel gives the highest accuracy result among three SVM models for lung cancer dataset. As for RCC dataset, it is found that the developed SVM with linear kernel gives the best classification accuracy among the SVM models. Therefore, evaluation of results from three different kernels types including linear, polynomial and RBF of SVM using RNA-Seq datasets yields that RBF gives more accurate results in binary classification and Linear kernel type performs the best in multi-class classification problems. It is found that the most of the developed deep learning models give better results compared to classical machine learning algorithms in terms of all metrics for classification of RNA-Seq cancer datasets (see Table 5-10). Moreover, Table 5 and Table 8 indicates that AdaDelta gives the best results for both datasets among the developed deep learning models with 7 different optimizers for all metrics. The developed model with AdaDelta has the lowest MAE values of 0.09 for lung cancer and 0.07 for RCC datasets. Thus, it is shown that more successful classification results can be obtained with the developed deep learning model using AdaDelta optimizer to classify RNA-Seq datasets.

In comparison of our results to another study available in literature [12] using different classification methods on the same datasets, it is found that the developed deep learning method with AdaDelta optimizer outperforms all methods applied in that study for both datasets. In that study [12], SVM and RF were the most accurate classifiers for RCC dataset with 93.5% and 92.3% test accuracy, respectively. As for lung cancer dataset, SVM and RF methods were again the most accurate classifiers with 94.8% and 93.8% test accuracy, respectively. In the present study, the developed deep learning model with AdaDelta optimizer yielded higher test accuracy results for RCC and lung cancer datasets with 96.15% and 95.54%, respectively. Finally, considering that the same datasets are used as input data for previous and the present research study, it should be noted that the developed deep learning model with AdaDelta has a significant improvement on the classification of the RNA-Seq cancer datasets.

In addition, different methods have been proposed in recent years for classifying microarray and RNA-seq data. For example, Rukhsar et al. [31] classified the RNA-Seq dataset of five cancer type using eight deep learning models. Highest accuracy was achieved with CNN of rates in the 90% and 95% among all methods. Haznedar et al. [32] proposed a hybrid method combining ANFIS and SA algorithm to classify five different microarray cancer data. The performance of the proposed method was compared with the classical ANFIS model and machine learning algorithms such as J48, BayesNet, and SVM. The obtained results showed that ANFIS-SA with an accuracy rate of 96.28% was the best performing model among all other methods. Khalifa et al. [33] classify five different RNA-seq data using hybrid model based on deep learning and BPSO-DT. The results showed that the overall accuracy of 96.90% was obtained with the proposed approach. Kim et al. [34] studied classification analyses of various microarrays to compare the performances of five classification algorithms over different data traits. It was observed that DT and RF achieved best performance with 95% accuracy rate and the remaining methods MLP, SVM, and KNN yielded worse performance with similar results for classification of lung cancer. In the present study, proposed DNN-Adadelta model achieved best accuracy rates of 95.54% and 96.15% for lung cancer and RCC RNA-seq datasets which are relatively close predictions compared to the recent studies above.

V. CONCLUSION

Cancer is one of the fatal diseases due to the fact that millions of people is dying because of or diagnosed with cancer, every year. Early diagnosis of cancer can be crucial for the success of the treatment and avoiding deaths due to cancer. Microarray and RNA-Seq technology provide gene expression of many genes simultaneously and help scientists understand which genes correspond to which cancer. The most effective 50 different genes are determined by applying wrapper feature selection method for renal cell cancer and lung cancer in this study. These findings based on machine learning and deep learning can be used to create a decision support system for doctors during the cancer diagnosis and classification purposes.

In this study, deep learning and classical machine learning algorithms’ performances are investigated for classification of RCC and lung cancer RNA-Seq datasets. In general, it is found that, the developed deep learning model are more successful than classical machine learning algorithms including DT, RF, SVM and ANN. Test accuracy, MAE and RMSE values are used for quantitative comparison of investigated models’ performances. As a result of comparisons based on test accuracy, MAE and RMSE values; it is found that the developed deep learning model with AdaDelta optimizer is the most successful in comparison to deep learning models with different optimizers, classical machine learning algorithms and simple artificial neural networks in classifying RNA-Seq cancer datasets. It is demonstrated that the developed deep learning model can be successfully used for analysis of RNA-Seq data for specific cancer types.

The optimization of the DNN model is one of the main problems in classification studies. The implementation of high performance optimization algorithm is needed to overcome such problem. For future prospects, optimization techniques based on artificial intelligence such as the Artificial Bee Colony (ABC), Particle Swarm Optimisation (PSO), Genetic Algorithm (GA), and Differential Evolution (DE) which are generally meta-heuristic algorithms may be used for the optimization of DNN model. Such optimization tech-
niques can be adopted to improve the performance of the present model.

REFERENCES

[1] A. Perez-Diez, A. Morgan, and N. Shulzenko, “Microarrays for cancer diagnosis and classification,” in Microarray Technology and Cancer Gene Profiling, S. Moccellin, Ed. New York, NY, USA: Springer, 2007, pp. 74–85.

[2] M.-S. Chua, E. Mansfield, and M. Sarwal, “Applications of microarrays to renal transplantation: Progress and possibilities,” Frontiers Biosci., vol. 8, no. 6, pp. 913–923, 2003.

[3] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, “Limma powers differential expression analyses for RNA-sequencing and microarray studies,” Nucleic Acids Res., vol. 43, no. 7, p. e47, Apr. 2015.

[4] Z. Wang, M. Gerstein, and M. Snyder, “RNA-seq: A revolutionary tool for transcriptionomics,” Nature Rev. Genet., vol. 10, no. 1, pp. 57–63, 2009.

[5] F. Ozsolak and P. M. Milos, “RNA sequencing: Advances, challenges and opportunities,” Nature Rev. Genet., vol. 12, no. 2, pp. 87–98, Feb. 2011.

[6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, Sep. 2015.

[7] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, “A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data,” Comput. Methods Programs Biomed., vol. 166, pp. 99–105, Nov. 2018.

[8] J. Huang, H. Fang, and X. Fan, “Decision forest for classification of gene expression data,” Comput. Biol. Med., vol. 40, no. 8, pp. 698–704, Aug. 2010.

[9] B. Han, L. Li, Y. Chen, L. Zhu, and Q. Dai, “A two step method to identify clinical outcome relevant genes with microarray data,” J. Biomed. Inform., vol. 44, no. 2, pp. 229–238, Apr. 2011.

[10] D. H. Tran, T. B. Ho, T. H. Pham, and K. Satou, “MicroRNA expression profiles for classification and analysis of tumor samples,” IEEE Trans. Inf. Syst., vol. 29, no. 3, pp. 416–422, 2011.

[11] H. Kim, G. H. Golub, and H. Park, “Missing value estimation for DNA microarray gene expression data: Local least squares imputation,” Bioinformatics, vol. 21, pp. 98–187, Jan. 2005.

[12] G. Zararsiz, D. Goksuluk, B. Klaus, S. Korkmaz, V. Eldem, E. Karabulut, and A. Ozturk, “VoomDDA: Discovery of diagnostic biomarkers and classification of RNA-seq data,” PeerJ, vol. 5, p. e3890, Oct. 2017.

[13] Y. Tan and P. Cahan, “SingleCellNet: A computational tool to classify single cell RNA-seq data across platforms and across species,” Cell Syst., vol. 9, pp. 207–213, Aug. 2019.

[14] J. Alquicira-Hernandez, A. Sathe, H. P. J. Li, Q. Nguyen, and J. E. Powell, “sCpred: Accurate supervised method for cell-type classification from single-cell RNA-seq data,” Genome Biol., vol. 20, no. 1, p. 264, Dec. 2019.

[15] N. Y. Simsek, “A comparative study of deep learning methods for classification of RNA-seq cancer data,” M.S. thesis, Inst. Natural Sci., Hasan Kalyoncu Univ., Gaziantep, Turkey, 2019.

[16] N. Y. Simsek, B. Haznedar, and C. Kuzudilisi, “A comparative study of different classification algorithms on RNA-Seq cancer data,” New Trends Issues Proc. Adv. Pure Appl. Sci., vol. 12, pp. 24–35, 2020, doi: 10.18844/gpjas.v12i2.4983.

[17] M. Saleem, S. S. Padmanabhan, A.-C.-N. Ngomo, J. S. Almeida, S. Decker, and H. F. Deus, “Linked cancer genome atlas database,” in Proc. 5th Int. Conf. Semantic Syst., 2013, pp. 129–134.

[18] R. Goyal, E. Gersbach, X. J. Yang, and S. M. Roohan, “Differential diagnosis of renal tumors with clear cytoplasm: Clinical relevance of renal tumor subclassification in the era of targeted therapies and personalized medicine,” Arch. Pathol. Lab. Med., vol. 137, no. 4, pp. 467–480, Apr. 2013.

[19] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” Neurocomputing, vol. 300, pp. 70–79, Jul. 2018.

[20] J. F. Tompson, A. Jain, Y. LeCun, and C. Bregler, “Joint training of a convolutional network and a graphical model for human pose estimation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 1799–1807.

[21] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep convolutional neural networks for LVCSR,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., May 2013, pp. 8614–8618.

[22] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, “Deep neural nets as a method for quantitative structure-activity relationships,” J. Chem. Inf. Model., vol. 55, no. 2, pp. 263–274, 2015.

[23] T. Ciodaro, D. Deva, J. M. de Seixas, and D. Damazio, “Online particle detection with neural networks based on topological calorimeter information,” J. Phys., Conf. Ser., vol. 368, no. 1, 2012, Art. no. 012030.

[24] M. Helmaoel, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, and W. Denk, “Connectomic reconstruction of the inner plexiform layer in the mouse retina,” Science, vol. 347, no. 6218, 2015, Art. no. 1254806.

[25] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Guerroussou, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, and B. J. Frey, “The human splicing code reveals new insights into the genetic determinants of disease,” Science, vol. 347, no. 6218, 2015, Art. no. 1254806.

[26] J. Deng and D. Yu, “Deep learning: Methods and applications,” Found. Trends Signal Process., vol. 7 nos. 3–4, pp. 197–387, 2014.

[27] Y. Bengio, “Learning deep architectures for AI,” Found. Trends Mach. Learn., vol. 2, no. 1, pp. 1–55, 2009.

[28] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.

[29] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature,” Geosci. Model Develop., vol. 7, no. 3, pp. 1247–1250, Jun. 2014.

[30] R. J. Hyndman and A. B. Koehler, “Another look at measures of forecast accuracy,” Int. J. Forecasting, vol. 22, no. 4, pp. 679–688, Oct. 2006.

[31] L. Rukhsar, W. H. Bangyal, M. S. Ali Khan, A. A. Ag Ibrahim, K. Nisar, and D. B. Rawat, “Analyzing RNA-seq gene expression data using deep learning approaches for cancer classification,” Appl. Sci., vol. 12, no. 4, p. 1850, Feb. 2022.

[32] B. Haznedar, M. T. Arslan, and A. Kalinli, “Optimizing ANFIS using simulated annealing algorithm for classification of microarray gene expression cancer data,” Med. Biol. Eng. Comput., vol. 59, no. 5, pp. 507–509, May 2021.

[33] N. E. M. Khalifa, M. H. N. Taha, D. E. Ali, A. Slowik, and A. E. Hassanien, “Artificial intelligence technique for gene expression by tumor RNA-seq data: A novel optimized deep learning approach,” IEEE Access, vol. 8, pp. 22874–22883, 2020.

[34] J. Kim, Y. Yoon, H.-J. Park, and Y.-H. Kim, “Comparative study of classification algorithms for various DNA microarray data,” Genes, vol. 13, no. 3, p. 494, Mar. 2022.