New Trends and Issues Proceedings on Advances in Pure and Applied Sciences

No 12 (2020) 24-35

www.propaas.eu

A comparative study of different classification algorithms on RNA-Seq cancer data

Nihat Yilmaz Simsek*, Computer Engineering Department, Hasan Kalyoncu University, Havaalanı Yolu Uzeri 8. Km, 27410 Sahinbey/Gaziantep, Turkey, https://orcid.org/0000-0003-0577-2766

Bulent Haznedar, Computer Engineering Department, Hasan Kalyoncu University, Havaalani Yolu Uzeri 8. Km, 27410 Sahinbey/Gaziantep, Turkey, https://orcid.org/0000-0003-0692-9921

Cihan Kuzudisli, Computer Engineering Department, Hasan Kalyoncu University, Havaalanı Yolu Uzeri 8. Km, 27410 Sahinbey/Gaziantep, Turkey, https://orcid.org/0000-0003-4774-152X

Suggested Citation:

Simsek, N. Y., Haznedar, B. & Kuzudisli, C., (2020). A comparative study of different classification algorithms on RNA-Seq cancer data. *New Trends and Issues Proceedings on Advances in Pure and Applied Sciences.* (12), 024–035.

Received date January15, 2020; revised date March 25, 2020; accepted date April 21, 2020.
Selection and peer review under responsibility of Prof. Dr. Dogan Ibrahim, Near East University, Cyprus.

Abstract

Gene mutations are the most important reason of cancer diseases, and there are different kind of causing genes across these diseases. RNA-Seq technology enables us to allow for gathering information about many genes simultaneously; hence, RNA-Seq data can be used for cancer diagnosis and classification. In this study, RNA-Seq dataset for renal cell cancer is analysed using three different developed classification methods: random forest (RF), artificial neural network (ANN) and deep learning (DL). The genes in our dataset are related to the following cancer types: kidney renal papillary cell, kidney renal clear cell and kidney chromophore carcinomas. It suggests that the DL method gives the highest accuracy rate compared to RF and ANN for 95.15%, 91.83% and 89.22%, respectively. We believe that the results acquired in this study will make a contribution to the classification of cancer types and support doctors in their processes of decision making.

Keywords: Classification, gene-expression, RNA-Seq, DL.

* ADDRESS FOR CORRESPONDENCE: **Nihat Yilmaz Simsek,** Computer Engineering Department, Hasan Kalyoncu University, Havaalani Yolu Uzeri 8. Km, 27410 Sahinbey/Gaziantep, Turkey. *E-mail address*: nyilmaz.simsek@hku.edu.tr / Tel.: +90-507-859-13-46

1. Introduction

Cancer is primarily a genetic disease which is widespread in our daily life. Generally, it 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 they can be occurred through somatic mutations. Diagnosis and classification of cancer by the gene expression has great 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 making a protein. By analysing these expressions, scientists can reach very useful information about the cancer. Microarray is one of the most well-known tools which used in the laboratory for detecting the expression of many genes at same moment. The data which is gathered from microarrays can be used for diagnosis and classification of human cancer [17].

More recent technology used for detecting 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 the RNA-Seq technology has started to become the major principle and commonly-used in gene-expression research studies [18]. [26] used the support vector machines (SVM) for the analysis of RNA-Seq data for detecting different cancer types. RNA-Seq data has a big dimensionality with many genes. During the diagnosis of cancer, most of the genes are not relevant. For example, in human genome, there are nearly 25,000 coding genes and 291 of them observed that caused to cancer [5]. This study showed that the number of genes can be minimised for using the classification or diagnosis of cancer. There are different methods for feature selection and one of them is wrapper method. In this study, wrapper method was used as feature selection method and applied on RNA-Seq dataset to reduce the number of genes and to find out the most effected genes for cancer diagnosis.

In the literature, several studies which have analysed gene expression datasets by using microarray and RNA-Seq technology. [10] applied three different classification algorithms to four different cancer microarray datasets. DT, SVM and *k*-nearest neighbours were applied as the classification algorithm, while hepatatox, colon cancer, lymph cancer and leukaemia microarray datasets were analysed. At the end of that study, DT method on leukaemia data yielded the highest accuracy rate of 96.6%. In another study, leukaemia, brain tumour, prostate and colon cancer datasets were analysed with four different classification algorithms [7]. Firstly, gene selections were made in datasets before classification algorithms were applied. Datasets were divided into subsets of 5, 10, 20, 50and 100 genes. Lastly, classification algorithms were applied and it was found that Naive Bayes algorithm achieved 91.1% accuracy rate on colon cancer. There also exist a few studies using RNA-Seq datasets in the literature because it is more recent compared to microarray technology. Tran, Ho, Pham & Satou (2011) have focused their work on microRNA (miRNA) data. Using the microarray data set used by [12], they classified samples as tumours and normal cells. This dataset has 223 samples with 151 miRNA properties. In that study, SVM with three different kernel types, including Linear, Polynomial and RBF were used on the dataset. 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. [25] applied 17different 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) gave the best accuracy. 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. [21] presented his own tool which named as Single Cell Net for the classification of single-cell RNA-Seq data. This tool compared favourably to other methods in sensitivity and specificity. [1], developed scPred as a new method can be used classification of RNA-Seq dataset using combination of unbiased feature selection and machine-learning based prediction method. A semi-supervised deep learning (DL) method used by [24] for the cancer prediction using RNA-Seq dataset. They developed a stacked sparce auto-encoder (SSAE) based method then SSAE tested on three different cancer RNA-Seq dataset, and it shown that the developed method has better classification performance in various metrics. In the present study, we have analysed the same RCC RNA-Seq dataset.

In the present study, machine learning algorithms including RF and ANN are selected and implemented for the purpose of comparison for their accuracies and performances to our proposed DL algorithm for the analysis of RCCRNA-Seq dataset. A feature selection method specifically wrapper method is applied on this dataset before applying the classification methods above.

2. Renal cell cancer (RCC) RNA-Seq cancer dataset

RCC which is an RNA-Seq dataset provided by the cancer genome atlas (TCGA) [20].There are many datasets for researchers to study, download and analyse in TCGA which is a comprehensive community resource platform. There are 1,020 RCC samples with 20,531 RNA transcripts for each sample in dataset which is taken from TCGA. 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 most known subtypes of RCC (account for nearly 90%–95% of the total malignant kidney tumours in adults) and separated as three different classes in this study [6].

Table 1. RCC dataset

Dataset	Number of samples	Number of genes	Number of Samples (each class)			Provided platform
RCC	1,020	20,531	KIRP	KIRC	KICH	The cancer genome atlas
			606	323	91	

3. Methods

In this section, feature selection methods, classification algorithms and evaluations methods of results are explained in detail.

3.1. Feature selection

When applying artificial intelligence to a problem, the most important process is to create a suitable model for the problem. One of the most important factors affecting the model's prediction process is the number of input elements. Feature selection is the process of decreasing the number of input elements. Feature selection is very important because the high number of input variables increases the calculation cost and time of the model and causes the performance to decrease. In this study, wrapper method was used to reduce the number of genes.

3.1.1. Wrapper method

Wrapper method is a feature selection method. It is used for creating last version of algorithm which will be used make a final classifier for feature subset selection. Therefore, A is a classifier and S is a feature set, then wrapper method looks for in the subset domain of S and trained classifier A is tested on each subset. Then, results are compared using by cross-validation method.

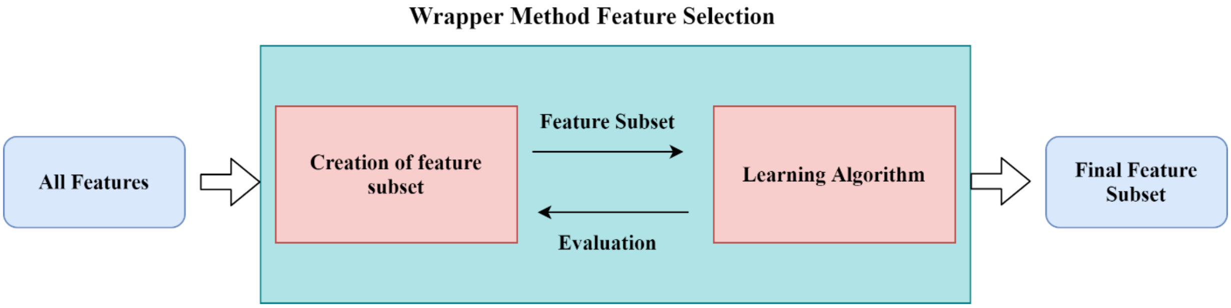

Figure 1. Wrapper method algorithm

Wrapper method is more computationally expensive than other feature selection techniques; however, it is better to have a good bias which is suitable for learning algorithm and it provides a better performance.

3.2. Classification

Classification is the process of separating the elements of a given data set according to its category. The classification process can be applied on both structured and unstructured datasets. The purpose of the classification is to find out which class a data is in. The class is generally called the target, tag or category. The classification model is used to find out which class the input data belongs to. The purpose of using the model here is to use it in the classification of new data that will come later. In this study, it is aimed to classify the data by using DL algorithms.

3.2.1. Random forest

The first studies about RF has been presented in the University of California by [3]. It is created from many other different and completely independent classifiers (decision tree). Given a test data as input to the new classifier can be classified according to the ranking of results from every different classification.

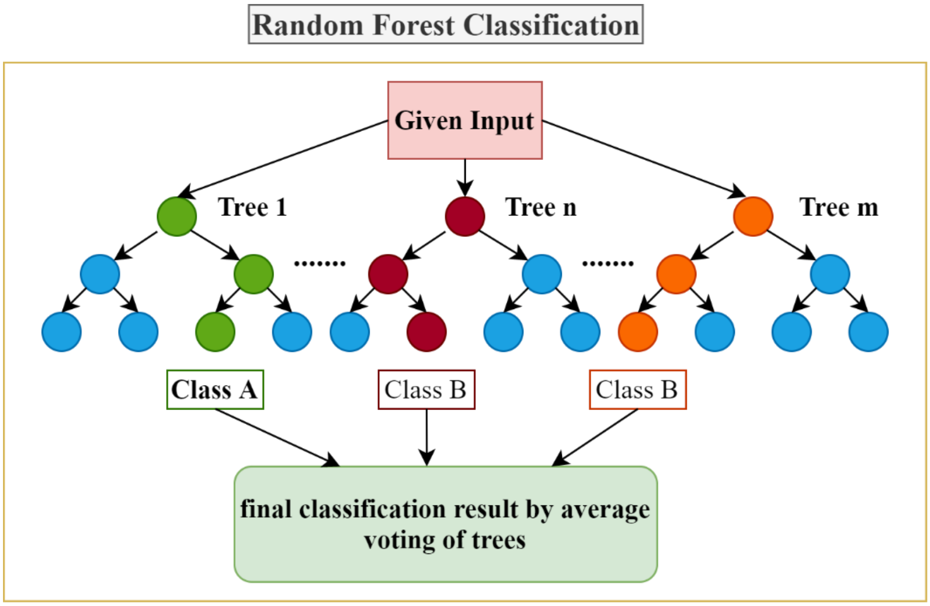
Random Forest Classification
Given Input
Tree 1
Tree n
Tree m
....
....
Class A
Class B
Class B
final classification result by average voting of trees

Figure 2. RF Algorithm

RF is occurred with a huge amount of decision trees. The randomness is the most important operation with choosing of examples subset and feature subset for creating a RF. It is very important to making independent decision tress, decreasing classification success and have better generalisation skills [3]. The random operation is used to have training subset from original examples with bagging method. This process is important for providing the independence of every preparation subset. The selected feature subsets are used as a training dataset. Rating of all features with respect to importance and results of every training results can affect the final decision. *N* variable is very critical for RF because of strengthless and correlation. Strengthless and correlation can be changed for a better result with the value of *N*. The advantage of random operation in RF is increasing the accuracy of classifier. It is very fast to create a single decision tree and RF uses the parallel use of these decision trees which decrease the classification time.

3.2.2. Artificial neural networks

Artificial neural networks (ANNs) are adaptive nonlinear data processing systems which merge many processing units with a series of features such as self-organising, self-adapting and real-time learning [16]. Studies on the ANNs have been significantly increased from 1980s and ANNs are applied to many problems in different areas. Many problems have occurred while studies on ANNs were increasing. For example, structure and parameter choice of the networks, dataset selection for training, stating the initial values are the some of these problems.

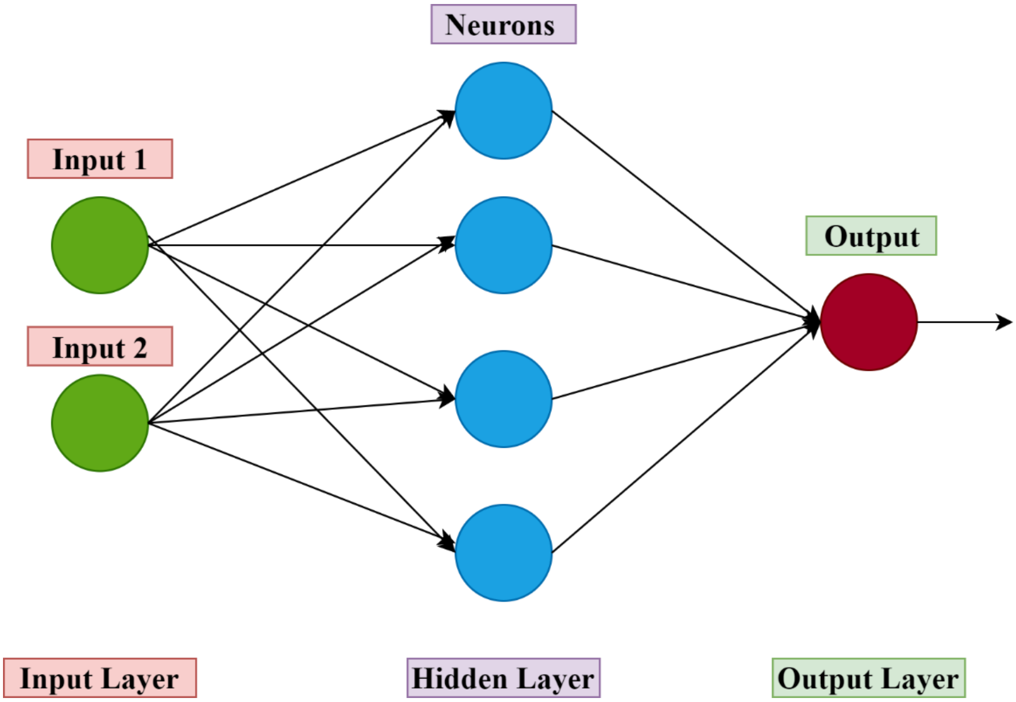

Figure 3. Structure of ANNs

The ANNS are used training dataset for learning process. It iteratively changes the values of weights to reach the desired output values. There are three main learning types in ANNs, these are supervised, unsupervised and reinforcement learning. The basic idea at behind of supervised learning is comparing the actual and desired result. Back propagation and other optimisation algorithms are used for decreasing error in result with iteratively adjust the weights. Reinforcement learning is separated from other supervised learning because it just checks actual output is correct or not. Finding the best correlation of the input data is the basic principal of unsupervised learning. There is just finding a rule for updating weights.

3.2.3. Deep learning

DL is a branch of machine learning and uses computational models which are formed of multiple processing layers to learn representations of data with high-level of abstraction. Very complex functions can be learned with sufficient combination of such transformations. For classification tasks, higher representation layers strengthen aspects of input that are important for discrimination and suppress irrelevant variations. One of the potentials of DL is changing manual features with effective algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction [13].

Despite the best suggestions of artificial intelligence, DL is making great progress in problems that cannot be solved for years. DL has seemed to be master at solving very complex problems and high-dimensional data for various fields such as science, business and government. Additionally, it has better results and performance than other artificial intelligence methods at image and speech recognition [19], [22] studies about drug molecules [15], analysing particle accelerator data [4], reconstruction of brain circuits [8] and prediction of mutation effects in non-coding DNA on gene expression and disease [23].

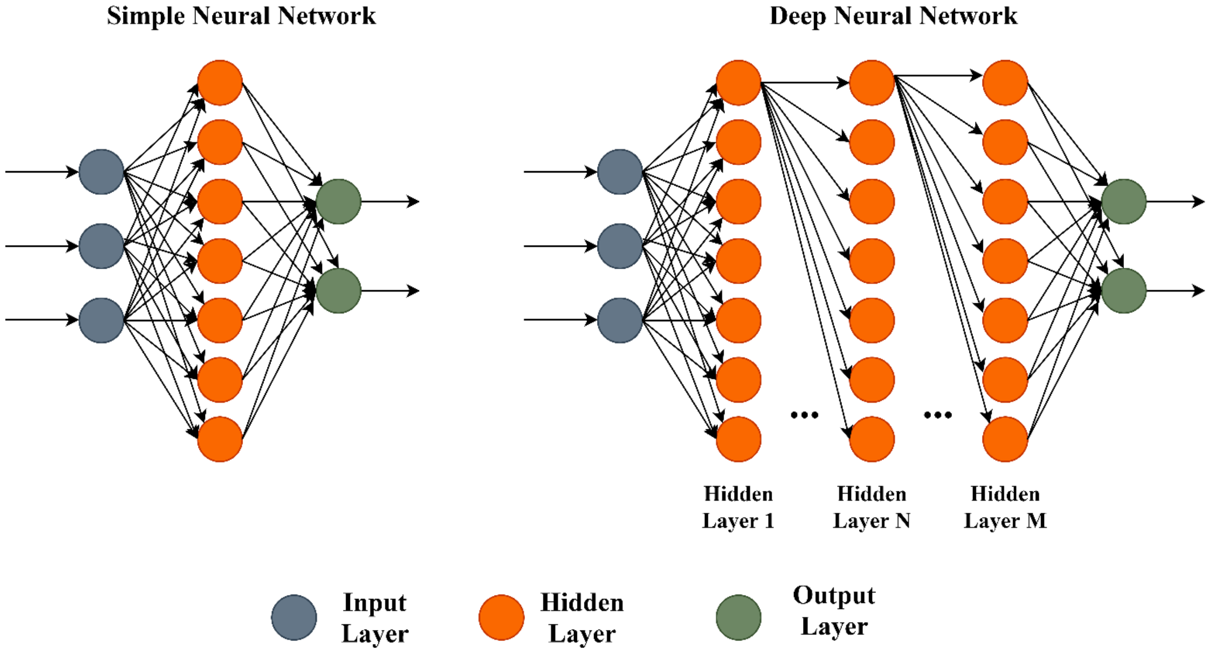

Figure 4. Neural networks vs DL architecture

There are different types of DL architectures, such as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks. In this study, a deep neural network model proposed with different optimisers. Next section gives the basic architecture of deep neural network.

3.3. Deep neural network architecture.

A Deep Neural Network (DNN) is in fact an artificial neural network (ANN) with several hidden layers of units across the input and output layers [14]. DNN can also get model of complex non-linear relationships like ANN. 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 [2]. DNN are generally aimed to function as feed forward networks and it can be discriminatively trained with the standard back-propagation algorithm. Stochastic Gradient Descent is used to update weights with the following Equation (1):

$$w_{ij}(t+1) = w_{ij}(t) + \mu \frac{\partial C}{\partial wij} \tag{1}$$

Where μ denotes the learning rate and C represents the cost function. The selection of the cost function is dependent 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, soft max function can be chosen as the activation function and cross entropy function can be used as the cost function. The soft max function can be described as

$$P_j = \frac{\exp{(x_j)}}{\sum_k exp\,(x_k)} \tag{2}$$

here, 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, respectively, of the same level. Cross entropy (cost function in supervised learning on multiclass classification problems) is formulated as

$$C_r = \sum_j d_j \, log \, (P_j) \tag{3}$$

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 [9].

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 learning a wide range of objects. The model comprises of multiple layers and each has a rectified linear unit for 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 in order to reduce L2 error to predict the mask ranging over the whole training set including bounding boxes represented as masks [9].

In this study, the model developed using these advantages of DL will be compared with other classical artificial intelligence methods.

3.4. Evaluation methods of classification results

3.4.1. Mean absolute error (MAE)

The MAE finds the average magnitude of errors in a series of estimates, regardless of their direction. It calculates the accuracy for continuous variables. The equation can be found in library references. The MAE is the average of the absolute values of the differences between the estimate and the coincident observation relative to the validation example. MAE is a linear score; this means that all individual differences are on average equal weight.

The MAE is given by:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n}$$

The MAE is an average absolute value of errors $|x_i - y_i|$, where x_i is the prediction and $y_i$is the target value.

3.4.2. Root mean squared error (RMSE)

RMSE is a quadratic scoring principle which also calculates the error's average magnitude. It is the square root of the mean differences in squares between prediction and real observation.

The MAE is given by:

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

The root mean square error is shown above, where x_i represents the prediction class and y_i represents the result of truly classified values. RMSD is a measure of truthiness that used for compare prediction errors of different models. It compares these errors not between datasets because it depends on the scale [11].

4. Results

In this section, wrapper method was used for the gene selection. There were 20,531 genes in one sample, and it was computationally very expensive to train it. Some of these genes were common in each sample and they had no effect on classification. Python programming language was used on Tensor Flow environment for applying a correlation method for genes. RF Regress or was used as the estimator in the training model. Dataset was divided into training set and test set. Eighty percent of dataset was used for training and 20% for testing. *K*-fold cross validation value was taken as 5 to increase the dataset variance. After running the program, the best suited 50 genes were selected for classification are shown in Table 2.

Table 2. Selected genes for RCC

ACBD7	CLDND2	LDLR	OBSCN	RGS22
ADAMTS18	CORO7	LOC283663	OR52W1	SCD
ADCY8	DENND1A	LOC285735	OTOP2	SNORD111B
C20orf96	EN2	LOC650293	PER4	SNX10
C2orf61	FABP7	MAP1B	PGBD1	SPIN2B
C2orf83|	GPR133	MCART6	PIP5KL1	TCEB3C
C6orf223	GPR144	MOS	PISD	TREML1
CCL7	HIST1H2BA	NACA2	PLAC1	WASF1
CCND2	KCNH4	NAMPT	PRSS42	WDR64
CCNO	KLB	NLRP10	PSMG1	ZFAT

After gene selection, classification algorithms were applied on RCC dataset. Firstly, classical algorithms were applied and then the results were obtained with DL method. Classical methods used were RF and ANNs. Afterwards, DL was applied and all these results were compared in Table 3.

As stated, the RF algorithm was applied firstly. Seventy percent of the dataset was reserved for training and 30% for the test. Initially, the number of trees was randomly assigned to 100. Then, the model was applied on the train set and then tested. As a result of the test, the model reached an accuracy rate of 91.83%.After classification, MAE value and RMSE were calculated. MAE value was calculated as 0.09 and 0.33 for RMSE.

After RF, the ANNs have been applied to RCC dataset. Similarly, dataset was split into 70%–30% for training and test, respectively. As a result of training and testing with ANNs, an accuracy rate of 89.22% was obtained. MAE value was calculated as 0.12 and 0.39 for RMSE.

Finally, our developed DL model was applied on RCC dataset. A sequential Keras model was used for DL model with three hidden layers. RMS Prop was selected as optimizer and Dropout was assigned as 0.5. Dataset was split into training and test. Seventy percent of our dataset was reserved for training and 30% for the test. Then, our model was trained and tested. Our DL model reached 95.15% accuracy rate with 0.07 MAE and 0.19 RMSE values. After application of these methods, results we represented in Table 3.

Table 3. Comparison of results for RCC dataset

Classifier	MAE	RMSE	Test accuracy
RF	0.09	0.33	91.83%
Artificial Neural Networks	0.12	0.39	89.22%
DL Model (RMSProp)	0.07	0.19	95.15%

Accuracy rates, MAE and RMSE of these three methods are shown in Figures4–6, respectively.

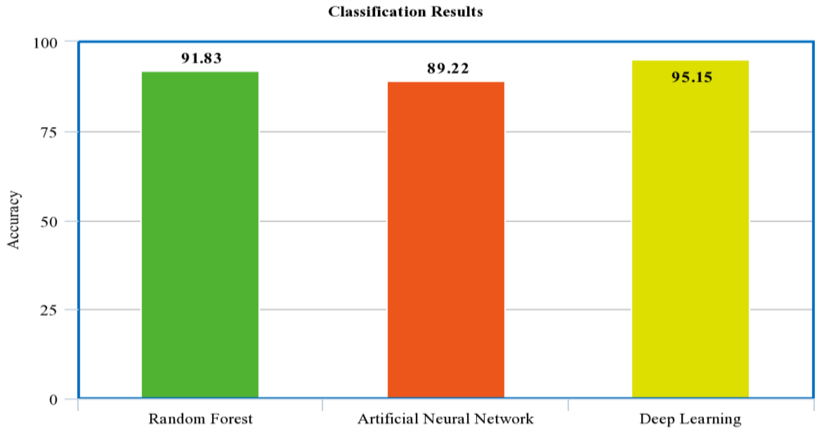

Figure 4. Accuracy results

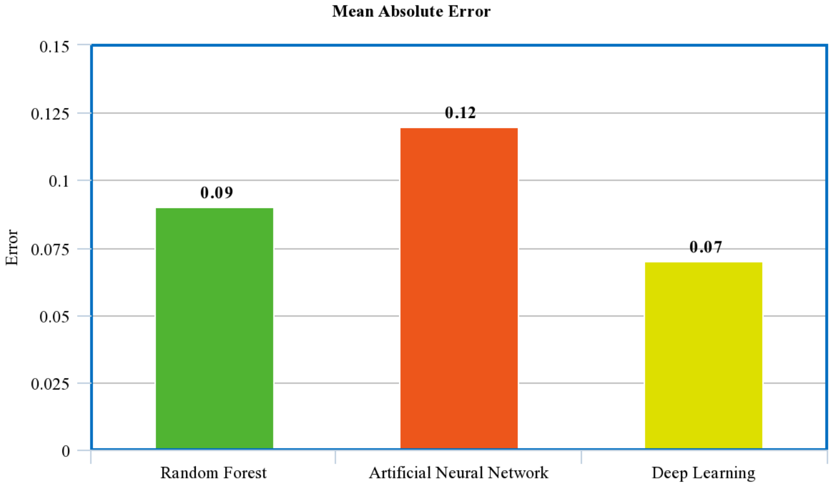

Figure 5. MAE results

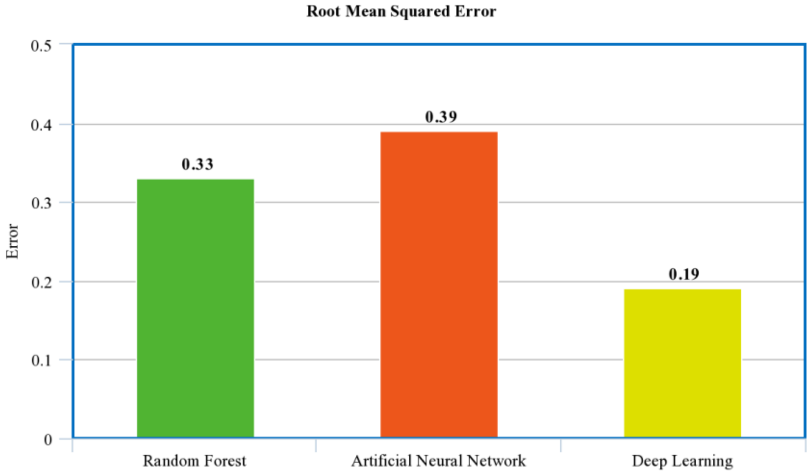

Figure 6. RMSE results

5. Discussions

In this study, the performance of DL and classical artificial intelligence algorithms' performances were compared by using RCC RNA-Seq dataset. It is shown that DL algorithms are more successful than RF and ANN for classification of RNA-Seq dataset. Three different evaluation criteria including test accuracy, MAE and RMSE used for the comparison of results and developed DL method reached the highest values. When results are compared with the study in the literature [1], [21], [22], [24], [25], developed DL method outperforms the all other methods in various metrics. In addition, applied feature selection method before classification shows that these 50 different genes are the most affective genes in human genome for RCC.

6. Conclusion

Cancer is one of the fatal diseases in our lives. Every year, millions of people are dying because of cancer and millions are diagnosed with cancer. In addition, time of cancer diagnosis plays a crucial role in the treatment. Microarray and RNA-Seq technology provide gene expression of many genes simultaneously and help us to understand which genes correspond to a disease. Therefore, RNA-Seq datasets can be used for diagnosis and classification of cancer diseases. These datasets can be used in machine learning and DL to create a decision support system for doctors during the cancer diagnosis and classification. In this study, we developed a DL model, ANN and RF model and then compared them.

Our results in Table 3 show that our DL model gave the best result among the classification algorithms applied after gene selection on RCC RNA-Seq data. While the DL model provided a success rate of 95.15%, the second closest result was obtained by RFs method with 91.83%. Training and test success rates can also be increased by using more datasets. Thus, the resulting reliability of the obtained system can be increased. In this study, RCCRNA-Seq dataset has been successfully used to make the decision support system. In conclusion, the cancer classification methods, which were proposed in this study, gave better results than previous studies. It is shown that these methods can be used for further analysis of RNA-Seq data for specific cancer types.

7. Recommendations

The RNA-Seq data we collected was limited for a comprehensive analysis of gene expressions and if more data is provided, a deeper insight into diseases classification could be gained. In addition, only wrapper method was used as feature selection in this study. However, there are alternative approaches such as Filter and Embedded methods and studying all these methods together could provide a better understanding of comparative results.

References

[1] Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. (2019). scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology, 20*, 264. doi:10.1186/s13059-019-1862-5

[2] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning, 2*, 1–127. doi:10.1561/2200000006

[3] Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. doi:10.1023/A:1010933404324

[4] Ciodaro, T., Deva, D., de Seixas, J. & Damazio, D. (2012). Online particle detection with neural networks based on topological calorimetry information. *Journal of Physics: Conference Series, 368*, 012030. doi:10.1088/1742-6596/368/1/012030

[5] Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer, 4*(3), 177–183. doi:10.1038/nrc1299

[6] Goyal, R, Gersbach, E., Yang, X. J. & Rohan, S. M. (2013). Differential diagnosis of renal tumors with clear cytoplasm. Clinical relevance of renal tumor sub classification in the era of targeted therapies and personalized medicine. *Archives of Pathology & Laboratory Medicine, 137*, 467–480.doi:10.5858/ arpa.2012-0085-RA

[7] Han, B., Li, L., Chen, Y., Zhu, L. & Dai, Q. (2011). A two-step method to identify clinical outcome relevant genes with microarray data. *Journal of Biomedical Informatics, 44*, 229–238. doi:10.1016/j.jbi.2010.11.007

[8] Helmstaedter, M., Briggman, K. L., Turaga, S. C., Jain, V., Seung, H. S. & Denk, W. (2013). Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature, 500*, 168–174. doi:10.1038/ nature12346

[9] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N. … Kingsbury, B.(2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *Signal Processing Magazine, IEEE, 29*, 82–97. doi:10.1109/MSP.2012.2205597

[10] Huang, J., Fang, H. & Fan, X. (2010). Decision forest for classification of gene expression data. *Computers in Biology and Medicine, 40*, 98–704. doi:10.1016/j.compbiomed.2010.06.004

[11] Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688. CiteSeerX 10.1.1.154.9771. doi:10.1016/j.ijforecast.2006.03.001.

[12] Kim, H., Golub, G. H. & Park, H. (2005) Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics, 21*(2), 187–198.doi:10.1093/bioinformatics/bth499

[13] LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444. doi:10.1038/nature14539

[14] Li,D. & Yu, D. (2014). Deep learning: methods and applications. *Foundations and Trends in Signal Processing,*
7(3–4), 197–387. doi:10.1561/2000000039

[15] Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure-activity relationships. *Journal of Chemical Information Modeling, 55*, 263–274. doi:10.1021/ci500747n

[16] McCulloch, W.S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133. doi:10.1007/BF02478259

[17] Perez-Diez A., Morgun A. & Shulzhenko N. (2007) Microarrays for cancer diagnosis and classification. InS. Mocellin (Ed.), *Microarray technology and cancer gene profiling. Advances in Experimental Medicine and Biology* (vol. 593). New York, NY:Springer.doi:10.1007/978-0-387-39978-2_8

[18] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. & Smyth, G. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research 43*(7), e47. doi:10.1093/nar/gkv007

[19] Sainath, T., Mohamed, A.-R., Kingsbury, B. & Ramabhadran, B. Deep convolutional neural networks for LVCSR. *Acoustics, Speech and Signal Processing,* 8614–8618. doi:10.1109/ICASSP.2013.6639347

[20] Saleem, M., Shanmukha, A., NgongaNgomo, A. C., Almeida, J. S., Decker, H. F. & Deus, H. F. (2013). *Linked cancer genome atlas database* (pp. 129–134). I-SEMANTICS ’13-Proceedings of the 9th international conference on semantic systems: 04–06 September 2013-Graz.doi:10.1145/2506182.2506200

[21] Tan, Y. & Cahan, P. (2018). SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. *Cell Systems, 9*(2), 207–213.e2. doi:10.1101/508085.

[22] Tompson, J., Jain, A., LeCun, Y. & Bregler, C. (2014). *Joint training of a convolutional network and a graphical model for human pose estimation*. Proceddings Advances in Neural Information Processing Systems, 27, 1799–1807.

[22] Tran, D. H., Ho, T. B., Pham, T. H. & Satou, K. (2011). MicroRNA expression profiles for classification and analysis of tumor samples. *IEICE Transactions on Information & Systems, 94*(3), 416–422. doi:10.1587/transinf.
E94.D.416

[23] Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C…Frey, B. J. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science, 347*, 6218. doi:10.1126/science.1254806

[24] Yawen, X., Jun, W. & Xiaodong, Z. (2018). A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data. *Computer Methods and Programs in Biomedicine, 166*. doi:10.1016/j.cmpb.2018.10.004

[25] Zararsiz, G., Goksuluk, D., Klaus, B., Korkmaz, S., Eldem, V., Karabulut, E., Ozturk, A. (2017). voomDDA: discovery of diagnostic biomarkers and classification of RNA-seq data. *PeerJ, 5*, e3890.doi:10.7717/peerj.3890.

[26] Zhang, Y. H., Huang, T., Chen, L., Hu, Y, Hu, L. D., …. Kong, X. (2017). Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets. *Oncotarget, 8*(50), 87494–87511. Published 2017 Sep 15. doi:10.18632/oncotarget.20903