Supervised Learning Approach for Knowledge Extraction & Decision-Making Process using Genome Sequence in Bioinformatics

May Abdullah Almutairi and Abdul Rauf Baig

College of Computer & Information Sciences,
Imam Mohammad Ibn Saud Islamic University (IMSIU),
Riyadh, Saudi Arabia.

Summary

Biomedicine, health care, and life sciences have recently played a significant role in data and information-intensive science. Particularly in the area of bioinformatics and computational biology, there is tremendous growth in data that could be noisy data, multidimensional, unstructured data or structured data, and the diversity of highly complex data. Therefore, a specific modelling and integrative analysis system is required. The present study focuses on developing a conceptual framework using deep learning approach to predictive modelling of diseases in bioinformatics using data from genome sequences. Initially, the data is pre-processed using Min-Max Standardization approach where it cross verifies the missing value and data scaling has been performed. Second, the significant features are selected using random forest method and it gets extracted using the deep learning-based auto-encoder method. Third, the data classification has been done with the help of XG-boost classifier technique. At last, the performance of suggested model has been tested using TCGA-PANCAN dataset then compared the performance with traditional method in terms of precision, recall, f-measure, accuracy, success rate, F-score and error rate.

Key words:
Genome Sequence, Machine Learning, Bioinformatics, Disease Prediction.

1. Introduction

Machine learning (ML) technique and computational intelligence (CI) is primarily applied to mining of data from huge set of data especially in the field of bio-molecular data [1], [2]. On the other hand, to predict and attain the some significant from bio-medical data via ML method. Here, ML prediction performed based on the behavior of neuron cell [3] that has been applied to bio-informatic data which differentiate the prokaryotic organisms then afterwards its used to various bio-informatics issues; for example proteomics, biological evolution, genomics, gene expression analysis, system biology, genomics and some other bio-informatics domains. Subsequently, ML and CI approach have also been applied to sequencing and reconstruction of genomes, identification and extraction of gene structures [4], [5], to the genome-wide identification of genes involved in genetic diseases [8], to identify RNA structural elements [9], analysis and identification of regulatory non-coding DNA elements [6], [7], to splice site prediction [11], to multiple alignments of bio-sequences in phylogenomics [13], to model haplotype blocks [10], to the detection and interactions of gene to gene of human diseases [12] and some other genomics issues. In the literature, some of the researcher suggested statistical and theoretical approach; however, the resulted output is not satisfactory with specific to measure of efficiency. Some of them stated that ML will be effective method for mining significant features via learning procedure that is applicable to huge and complex high dimensional data sets [14]. Even for this method, the disease prediction of genome sequence is becomes more complex [14]. For resolving this issue, there a need of effective ML method that require to predict and analyze the complex the data with high precision value [15]. A study by Sindhu [17] recommended hybrid method namely soft computing and data mining techniques by predicting or recognizing human diseases whereas it could be effective toward attain data in huge volume of data [16].

Especially, supervised ML method is effective method for disease prediction [18]. So, the present study relies on a supervised learning-based approach for data analysis of genomic sequence and predicts trends of diseases, particularly cancer disease with the highest precision of classification.

This study has been well organized into six section. Especially section 2 presents the concepts and theory background of ML in bioinformatics. Section 3 describes the proposed research methodology, whereas it summarizes the input data source, data pre-processing, feature selection, extraction, classification methods, and evaluation metric of genome sequence. Section 4 presents the simulation results, average measure of error rate and accuracy measurement. Section 5 summarize the obtained results and compared with traditional method and concluded in section 6.

Manuscript received December 5, 2019
Manuscript revised December 20, 2019
https://doi.org/10.22937/IJCSNS.2020.20.12.16

2. Machine learning in bioinformatics

Generally, ML approach learn the data pattern or rules of complex data in automatic manner especially for prediction and data representation problems, instead of explicitly defining them on the basis of prior domain observation or knowledge extraction. Even though a dramatically expanding number of AI procedures have been applied to take care of testing related issues, the trouble of model understanding is a hindrance that keeps on causing delay in using AI in specific zones. Moreover, the ongoing appearance of profound realizing, which is considered as a discovery, makes the trouble all the more testing, regardless of its exceptional prescient presentation in numerous applications [19]. A Machine learning calculation is seen as a subset of man-made brainpower where the consistent examination of calculations and real models are utilized. It successfully achieves a specific undertaking relying upon examples and derivation without using express rules. AI calculations use preparing information for settling on choices without being modified to complete the undertaking [20].

In the post-genome period, when the size of genomic information keeps on extending, specialists are confronting the challenges to oversee and comprehend the immense measure of information [21]. Computational calculations are being created to utilize genomic grouping information while the greater part of the information examination centers around overpowered measurements [22]. In the course of recent years, a progression of AI (ML) calculations have been created, and effectively applied in the field of bioinformatics [23], [24]. One effective use of ML on bioinformatics is quality capacity explanation which allocates quality philosophy terms to unannotated qualities [25], [26]. Moreover, ML assumes a critical function in uncovering quality collaborations [27], for instance, multifactor dimensionality reduction (MDR) utilize ML ways to deal with identify high request hereditary cooperations [28]. ML has likewise been applied in populace hereditary qualities, different ML calculations have been produced for the induction of segment narratives, populace size, recombination rates and examples of populace parting and movement [29]. Random forest (RF) calculation, which is an alleged gathering tree calculation, was utilized for disease arrangement just as tumor biomarker distinguishing proof [30]. This is of incredible criticalness for the early discovery and treatment of tumors. At present, improvement of ML examination techniques in bioinformatics become speedier and more advantageous, because of the accessibility of universally useful ML libraries including Scikit-learn, TensorFlow and Keras et al. It gives better occasions to non-PC bio-scientists to have the option to effectively deal with organic information and find new information.

3. Research methodology

In the present study, we have assumed three machine learning technique namely, deep learning-based autoencoder, random forest method and XGBoost classifier method. Here the deep learning technology is used for the reduction of dimensionality i.e extraction of significant features. Then the features were selected via random forest method and XGBoost classifier is used to classify the cancer type. In our research, the classifier XGBoost correctly identified the type of cancer of BRCA, COAD, KIRC, LUAD and PRAD. Finally, the performance of the suggested model has been tested using TCGA-PANCAN dataset then compared the performance with traditional method in terms of precision, recall, f-measure, accuracy, success rate, F-score and error rate. The simulation has been done with the system configuration of Intel(R) Core (TM) i7 processor with 16 GB of RAM running 64-bit Windows 10 Operating System and Deep Learning libraries based on Python programming. The conduction of this study as,

- First, a real word and publicly accessible data sets pertaining to cancer disease (genome sequence data from TCGA dataset) is used and to predict the disease status using the genomic sequence.
- Second, we adopted a Random Forest Classifier (RFC) approach for features selection.
- Third, we applied Deep learning based Autoencoders with hierarchical learning processes to forecasting the clinical events with respect to their internal validity and precision, where it will improve patient care.
- At last, we have validated the results and compared the results with traditional method in terms of accuracy, precision, FP rate, Recall, F1 score, success rate and error rate.

The implementation process flow of the proposed disease predictive scheme is illustrated in figure 1.

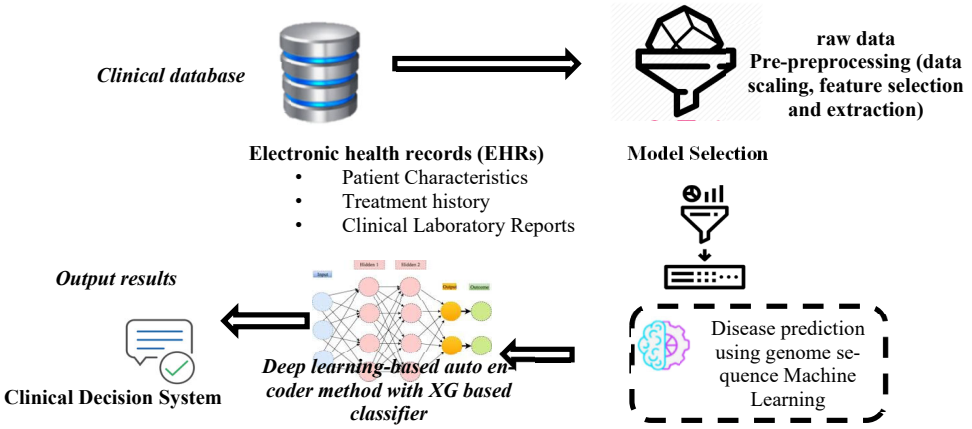

Figure 1: Proposed system architecture

3.1 Data source

The proposed predictive model performance has been tested using TCGA-PANCAN dataset. The input data as Cancer Genome Atlas (TCGA) program is a collection of clinico-pathological annotation data with multi-platform molecular profiles of over 11,000 human tumours from 33 different types of cancer. This data is compatible with cancer genomics research independent of the TCGA program and provides incentives to use clinical comparisons for unparalleled scale analysis of cancer biology. The TCGA-PANCAN dataset consists of 801 instances (rows) and 20531 attributes (columns) [31]. Further, in the dataset, the instances (samples) are stored in row-wise. The attributes that are variables of each sample are the RNA-Seq levels of gene expression calculated by the illuminated HiSeq platform. The partial view of dataset is represented in Table 1.

Table 1: Partial view of the data set

Samples/gene	gene_0	gene_1	gene_2	gene_3	gene_4	gene_5	gene_6	gene_7	gene_8	gene_9	gene_10	gene_11	gene_12
sample_0	0	2.017209	3.265527	5.478487	10.432	0	7.175175	0.591871	0	0	0.591871	1.334282	2.015391
sample_1	0	0.592732	1.588421	7.586157	9.623011	0	6.816049	0	0	0	0	0.587845	2.466601
sample_2	0	3.511759	4.327199	6.881787	9.87073	0	6.97213	0.452595	0	0	0	0.452595	1.981122
sample_3	0	3.663618	4.507649	6.659068	10.19618	0	7.843375	0.434882	0	0	0	0.434882	2.874246
sample_4	0	2.655741	2.821547	6.539454	9.738265	0	6.566967	0.360982	0	0	0	1.275841	2.141204
sample_5	0	3.467853	3.581918	6.620243	9.706829	0	7.75851	0	0	0	0.51541	0.51541	2.516797
sample_6	0	1.224966	1.691177	6.572007	9.640511	0	6.754888	0.531868	0	0	3.173927	1.476796	3.023841
sample_7	0	2.854853	1.750478	7.22672	9.758691	0	5.952103	0	0	0	0.441802	0	2.405856
sample_8	0	3.992125	2.77273	6.546692	10.48825	0	7.690222	0.352307	0	4.067604	1.411318	1.252839	2.579977
sample_9	0	3.642494	4.423558	6.849511	9.464466	0	7.947216	0.724214	0	0	0	1.204141	2.296311
sample_10	0	3.492071	3.553373	7.151707	10.25345	0	8.301258	0	0	0	0	1.999567	3.381962
sample_11	0	2.941181	2.663276	6.56169	9.376293	0	7.860323	0.754118	0	0	2.449641	1.021409	3.153092
sample_12	0	3.970348	2.364292	7.145443	9.240605	0	7.810758	0	0	0	1.12201	1.565987	2.698263

3.2 Data Pre-processing

In our research, the clinical data is pre-processed in four phases such as checking for missing value, Feature scaling, Feature selection and feature extraction. The detailed explanation is given as follows.

3.2.1 Checking for missing value

The first step of pre-processing is to check for the missing value in the dataset. We have analysed the TCGA-PANCAN dataset. As a result, we found that the dataset does not contain null elements. All values in the dataset are numerical values.

3.2.2 Feature scaling (Normalization)

Once the dataset has been checked with the missing values. It is processed for the next preprocessing step called feature scaling. The feature scaling is also called as data scaling. The data needs to be scaled before modelling. In our research, data scaling is carried out using the Normalization technique. All the data points in the TCGA-PANCAN dataset are scaled using the Min-max Normalization technique [32]. The Min-max Normalization technique will change the distribution shape of the data [33]. The numerical range value of the data feature is converted to the lower scale and fit between 0 and 1. The following formula is utilized to estimate the normalization z,

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Where,
Z – Normalization
x - Set of observed value
max(x) – Maximum value
min(x) – Minimum Value

3.2.3 Feature Selection and extraction approach

The feature selection is a filter process, used to pick the most important element from the dataset. The feature selection is used to improve accuracy, reduce overfitting and to reduce training time [34]. In this study, the random forest algorithm is used for feature selection and identifies the important features automatically from the dataset. The tree-based technique utilized in the random forest algorithm will rank and enhance the purity of the node [35]. Then the impurity is decreased from the tree using the mean technique called gini impurity. In addition, random forest technique takes only a small subset of features rather than all features. Further, in the mathematical theory of communication, the concept of information theory is used by the random forest method to pick the most significant feature by looking into a prediction variable.
The feature extraction is also known as dimension reduction. The feature extraction is a method to reduce dimensionality by reducing the original set of raw information to more workable processing groups. Here for the dimension reduction, the Deep learning autoencoder technique is used [36]. Autoencoders are a specific type of neural network structures in which the output is identical to the input [37]. In order to learn the incredibly low-level interpretations of the input data, autoencoders are educated or trained in an unsupervised way. The selected features from the random forest are then transferred to the neural model auto-encoder to minimize the dimension. The 205 attributes feature from the TCGA-PANCAN dataset is reduced to 12 principal features using the neural model auto-encoder technique. Figure 2 shows 5 labels that can be linearly separated with features 12 reduced dimensions.

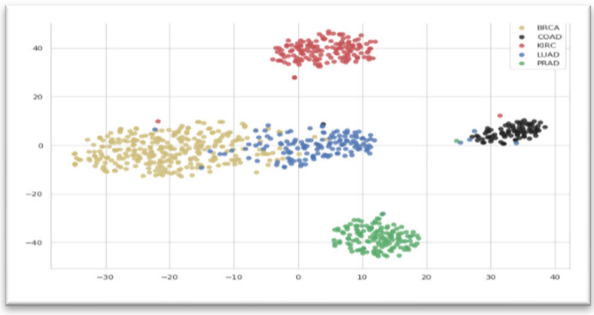

Figure 2: Feature Extraction

3.3 XGBoost classifier method

Once the data is pre-processed, it is necessary to train the data to accurately predict the results. For training the data, we need a supervised machine learning classification algorithm. In our research, we have used XGBoost classification algorithms to train the data [38]. Once the data is trained, learnt data is sent to the testing phase; the splitting ratio of TCGA-PANCAN dataset for training is 80%, and testing is 20%. The classifier predicts, where the patient has BReast CArcinoma (BRCA), COlon Adenocarcinoma (COAD), KIdney Renal Clear-cell carcinoma (KIRC), LUng ADenocarcinoma (LUAD), or PRostate Adenocarcinoma (PRAD). In brief, 640 features are used for training and 161 features used for testing. Detailing of cancer classes are given as follows, BRCA cancer type used 238 features for training and 62 features were used for testing. Similarly, for COAD cancer type, 69 features were used for training and nine features used for testing. Whereas for KIRC cancer type, 106 used for training and 40 features used for testing. Further, for

LUAD cancer type 121 used for training and 20 used for testing. Furthermore, for PRAD cancer type, 106 used for testing and 30 used for testing. Table 2 represents the split training and testing data ratios in detail.

Table 2: Split data in a ratio of 80% for training and 20% for testing

Cancer classes	Train count	Test count
BRCA	238	62
COAD	69	9
KIRC	106	40
LUAD	121	20
PRAD	106	30
TOTAL	640	161

3.4 Performance evaluation metric

The performance has been evaluated using precision, recall, f-score, success rate, error rate, accuracy. The resulted outcome has been compared with the traditional method (artificial neural network, decision tree and Bayesian approach). The term precision is measured by average number of positive predictions divided by the overall number of predicted positive class values. Here the low precision may also indicate as a great number of false positives. The worst is 0.0, whereas the best precision is 1.0. The mathematical representation of precision score is also defined as,

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN}+\text{FP}} = 1 - \text{SP}$$

$$\text{ERR} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}$$

Where,
ACC – Accuracy; TP – True Positive
TN – True Negative; FP – False Positive
FN – False Negative; Prec - Precision
REC – Recall; ACC- accuracy
FPR- false positive rate; SP – Specificity
ERR- Error rate.

The term recall (REC) clarifies that the model sensitivity in the way of finding the positive class. It is also called as the True Positive Rate (TPR) or Sensitivity. It evaluated as the ratio of true Positive to the total amount of true positives and false negatives. Recall is viewed as an indicator of completeness of the classifiers. Most False Negatives suggest a weak recall. The best recall is 1.0, while the worst is 0.0. The mathematical representation of recall is written as,

The F-measure, which is the harmonic mean of recall and precision, is also known as the F1-score. The range for the value of F-Measure is from 0 to 1. The high score is reflected by F Measure's high value. The F1 Score is also known as the F Rating or the F Index. The F1 score, to put it another way, expresses the balance between precision and the recall. The F-measure formula is given below.

$$\text{F1} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where,
F1 – F- measure
Accuracy (ACC) is measured as the number of accurate predictions divided by the total dataset number [39]. Accuracy identifies the positive classes and negative classes of the model. It calculated as the ratio of the total of real positive and real negative to entire samples (true positive and false positive, true negative and false negative). The accuracy rate is signified as closed to their real output. The best accuracy is 1.0, while the lowest is 0.0. It can be determined by 1-ERR as well. The Accuracy formula is given below.
The False Positive Rate (FPR) is measured as the number of wrong positive predictions divided by the total negative number. With respect to the proposed model, the false positive rate (FPR) is the number of people without the disease but reported as having the disease (all Positive), divided by the total number of people without the disease (including both false positive and True negative). It could also be measured as 1 – specificity. The worst false positive rate is 0.0, and the best false positive rate is 1.0. The False Positive Rate (FPR) formula is given below.
Error rate (ERR) is the sum of all incorrect predictions divided by the total number of the data. It is measured as two incorrect predictions of total number of disease (False Negative and False Positive) divided by total number of a data. One of the most common and logical measures extracted from the confusion matrix are error rate (ERR) and other is Accuracy (ACC). The best error rate is 0.0, and the worst is 1.0. The Error rate (ERR)) formula is given below.

Table 3: Connotation in terms of our predictive model

Term	Meaning
TP	A person has diseases, and the model correctly predicted that a person has the disease
TN	A person does not have diseases, and the model correctly predicted that a person doesn't have the disease
FP	A person does not have diseases and the model wrongly predicted has that person have the disease
FN	A person has diseases and the model wrongly predicted has that person does not have the disease

The confusion matrix is a good preference to report results in n-class classification issues because the relationship between the classifier outputs and the true ones can be observed. In other words, the confusion matrix provides a matrix as output and defines the model's full performance. Using confusion matrix, the True Positive (TP), True negative

(TN), False positive (FP) and False negative (FN) were calculated. Table 4 provides the detailed description of the confusion matrix and the actions carried out against each technique.

Table 4: Confusion Matrix

CLASS	Y	N
Y	True positive (TP)	False negative (FN)
N	Fasle positive (FN)	True negative (TN)

1. TP: Classified or identified correctly.
2. FP: Classified or recognized incorrectly. It reflects the error type I.
3. FN: wrongly ignored. It reflects the error type II.
4. TN: correctly ignored.

4. Experimental setup and Results

In this research work, an effective disease predictive modelling in medical application is implemented in python using four machine learning techniques - Min - Max Standardization, Random Forest, Deep learning autoencoder and XGBoost classifier. The study's experimental findings were all performed on a computer with a high visual interface configuration and operating system setup. The experiment is conducted on the test machine configured with Intel (R) Core (TM) i7 processor with 16GB of RAM running 64-bit Windows 10 Operating system. The data transformation and model training was executed using python 3.7 software. The detail of the system configuration is represented in table 5.

Table 5: System Configuration

System Specifications	Configuration Details
System Type	64 bit, Windows 10 Operating system
Processor Name	Intel (R) Core (TM) i7
RAM	16 RAM
Python	3.7

While a confusion matrix contains all information of the outcome of a classifier, they are rarely used for reporting results in Brain-Computer Interfaces (BCI) field because they are difficult to compare and discuss. Instead, certain parameters are usually removed from the confusion matrix. Below is a valuable table that provides a summary of error forms in both the class distribution in the data and the classifiers expected class distribution. Below table 6 provides the detailed description of the confusion matrix and the actions carried out against each cancer.

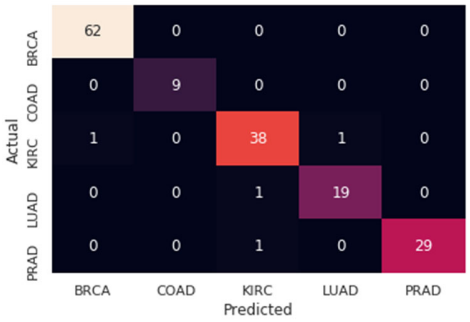

Figure 3: The predictive model evaluation using the Confusion matrix

The five cancer types such as BReast CArcinoma (BRCA), COlon Adenocarcinoma (COAD), KIdney Renal Clear-cell carcinoma (KIRC), LUng ADenocarcinoma (LUAD), and PRostate Adenocarcinoma (PRAD) were evaluated and predicted using a confusion matrix. The Number of data used for testing the cancer type such as BRCA, COAD, KIRC, LUAD and PRAD is 62, 9, 40, 20 and 30, respectively. The details of all cancer type test data are explained in chapter 3. Here in terms of the confusion matrix, our predictive model predicted 62 number of BRCA cancer type, it means it predicted 100% correctly. Similarly, for COAD cancer type, true positive is 9. It means for COAD cancer type also our predictive model predicted 100% correctly. Whereas for KIRC, LUAD and PRAD cancer type, Prediction of true positive is 38,19 and 29. Our model has predicted almost 95% correctly. The detailed portrayal is given in table 6.

Table 6: Actual vs Prediction values

		Predicted				
		BRCA	COAD	KIRC	LUAD	PRAD
Actual	BRCA	62	0	0	0	0
	COAD	0	9	0	0	0
	KIRC	1	0	38	1	0
	LUAD	0	0	1	19	0
	PRAD	0	0	1	0	29

The performance of the proposed system is examined by contemplating the actual and predicted classification. The framework suggested consists of three approaches such as

random forest, deep learning autoencoder and XGBoost classifier. All three approaches are performed to obtain the performance metric. The Accuracy, Precision, Recall, F1 score, Success rate and Error rate are calculated in the performance metric. Below is a detailed description of the performance metric and the actions carried out against each confusion matrix. Figure 4 represents the performance evaluation for the proposed model.

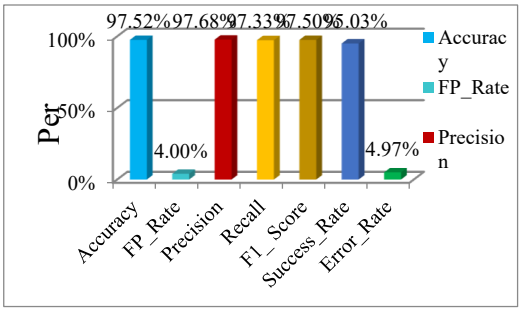

Figure 4: Performance Metric

While comparing the performance metric for the predictive model, we achieved 97.52 % of accuracy, 4.0% of FP_ rate, 97.68% of precision, 97.33% of recall, 97.5% of F1-score, 95.03% of success rate and 4.97% of Error_rate. The performance in term of accuracy, precision, recall and F1-score is efficient. Error_rate and FP_rate are less. Hence, we can conclude the predictive model is efficient. Now, let us compare genome sequence in Bioinformatics in term of the accuracy of the proposed technique with the previous research technique. Table 7 represents the comparison table.

Table 7: Comparison Table of Predicit Model Vs Existing Model

S.no	Author	Accuracy
1.	Piecemeal & Adenoma,(2007)	92%
2.	Zhu et al., (2020)	85.6%
3.	Proposed Model	97.52 %

The performance metric accuracy result of genome sequence in Bioinformatics disease is compared with different existing machine learning technique. Where, Zhu et al., (2020) obtained an accuracy of 85.6% whereas Piecemeal & Adenoma, (2007) predicted accuracy of 92% and our proposed model predicted accuracy of 97.52%. We found that our proposed model provides a better result when compared to other studies, accuracy results. Figure 5 represents the performance comparison chart for the proposed model vs existing model.

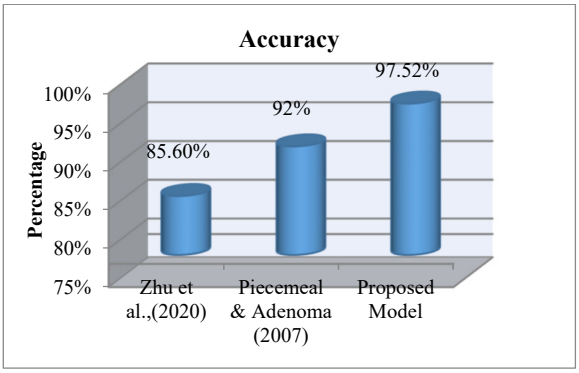

Figure 5: The performance comparison chart for the proposed model vs existing model

5. Discussion

Once the TCGA-PANCAN cancer dataset is trained and tested, and it is sent to the evaluation of the metrics. The metric is evaluated using a confusion matrix and performance metric. Using the confusion matrix values such as true positive, true negative, false positive and false negative were calculated. Whereas, using the performance metric, the accuracy, FP_rate, precision, recall, F1 score, success rate and the error rate were calculated. The output performance of the proposed model is evaluated using the actual classification value and predicted classification value.
False positives are no recurrence that was defined by the classifier as recurrence. False negatives which are labelled as no recurrence by the classifier are recurrence. The confusion matrix is a good choice to disclose results in classification issues of n-class since it is possible to observe the relationship between the classifier outputs and the true ones. The confusion matrix, in other words, provides a matrix as output and determines the maximum performance of the model. The True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) were determined using a confusion matrix. The confusion matric result is evaluated using actual and prediction values. The actual data used to test types of cancer such as BRCA, COAD, KIRC, LUAD, and PRAD are 62, 9, 40, 20, and 30. Our predictive model predicts data to be 62, 9, 38, 19, 29, respectively. The model effectively predicted the cancer type.

The performance metric is determined by enforcing actions against each confusion matrix. The performance of the proposed system is examined by contemplating the actual and predicted classification. The Accuracy, Precision, Recall, F1 score, Success rate and Error rate are calculated using the performance metric. Comparing the performance metric for the predictive model, we achieved 97.52 percent accuracy, 4.0 percent FP rate, 97.68 percent accuracy, 97.33 percent recall, 97.5 percent F1-score, 95.03 percent success

rate and 4.97 percent error rate. The performance is efficient in terms of precision, accuracy, recall, and F1-score. Error rate and FP rate are lower. While comparing the performance metric for the predictive model, we achieved 97.52 % of accuracy, 4.0% of FP_ rate, 97.68% of precision, 97.33% of recall, 97.5% of F1-score, 95.03% of success rate and 4.97% of Error_rate. The performance in term of accuracy, precision, recall and F1-score is efficient. Thus, we can conclude that the predictive model is accurate.

6. Concluding remarks

In this paper, we presented three different algorithms that have been developed for TCGA-PANCAN cancer dataset. We have especially applied random forest classifier for feature selection and identifies the important features automatically from the dataset. Subsequently, the feature extraction method was applied the to reduce dimensionality of genomic data by reducing the original set of raw information to more workable processing groups. Here for the dimension reduction has been done via deep learning autoencoder technique. Also, we have extracted the significant features and classified the types of cancer such as BRCA, COAD, KIRC, LUAD, and PRAD using genomic sequence. The outcome has been compared with the traditional method (genome-scale metabolic model, artificial intelligence) in terms of sensitivity, specificity, accuracy, f-score, success rate and error rate. Simulations have been performed using Python programming language. The model is trained on a significantly large amount of data and we have assumed different parameters. This model is also capable to overcome the data overfitting and then to import the significant extracted information to deep learning training phase. So, there is a better chance of generalization which keeps the model stable. The model ends up with the accuracy of 97.52%, 4.0% of FP_ rate, 95.03% of success rate and 4.97% of Error rate and an average precision, recall, specificity and f1-score as are 97.68%, 97.33% and 97.5% respectively.

References

[1] D. C. Wallace and D. Chalkia, “Mitochondrial DNA Genetics and the Heteroplasmy Conundrum in Evolution and Disease,” Cold Spring Harb. Perspect. Biol., vol. 5, no. 11, pp. a021220–a021220, Nov. 2013.

[2] E. M. McCormick, Z. Zolkipli-Cunningham, and M. J. Falk, “Mitochondrial disease genetics update,” Curr. Opin. Pediatr., vol. 30, no. 6, pp. 714–724, Dec. 2018.

[3] D. C. Wallace, “Mitochondrial DNA Variation in Human Radiation and Disease,” Cell, vol. 163, no. 1, pp. 33–38, Sep. 2015.

[4] M. R. Brent and R. Guigo, “Recent advances in gene structure prediction,” Curr. Opin. Struct. Biol., vol. 14, no. 3, pp. 264–272, 2004.

[5] A. Bernal, K. Crammer, A. Hatzigeorgiou, and F. Pereira, “Global discriminative learning for higher-accuracy computational gene prediction,” PLoS Comput Biol, vol. 3, no. 3, p. e54, 2007.

[6] G. RAtsch et al., “Improving the Caenorhabditis elegans genome annotation using machine learning,” PLoS Comput Biol, vol. 3, no. 2, p. e20, 2007.

[7] D. T. Holloway, M. Kon, and C. DeLisi, “Machine learning for regulatory analysis and transcription factor target prediction in yeast,” Syst. Synth. Biol., vol. 1, no. 1, pp. 25–46, 2007.

[8] N. Lopez Bigas and C. A. Ouzounis, “Genome wide identification of genes likely to be involved in human genetic disease,” Nucleic Acids Res., vol. 32, no. 10, pp. 3108–3114, 2004.

[9] L. Bao and Y. Cui, “Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information,” Bioinformatics, vol. 21, no. 10, pp. 2185–2190, 2005.

[10] G. Greenspan and D. Geiger, “High density linkage disequilibrium mapping using models of haplotype block variation,” Bioinformatics, vol. 20, no. suppl_1, pp. i137–i144, 2004.

[11] Y. Saeys, S. Degroeve, D. Aeyels, P. Rouzé, and Y. Van de Peer, “Feature selection for splice site prediction: a new method using EDA-based feature ranking,” BMC Bioinformatics, vol. 5, no. 1, p. 64, 2004.

[12] M. D. Ritchie, B. C. White, J. S. Parker, L. W. Hahn, and J. H. Moore, “Optimizationof neural network architecture using genetic programming improvesdetection and modeling of gene-gene interactions in studies of humandiseases,” BMC Bioinformatics, vol. 4, no. 1, p. 28, 2003.

[13] J. Handl, D. B. Kell, and J. Knowles, “Multiobjective optimization in bioinformatics and computational biology,” IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 4, no. 2, pp. 279–292, 2007.

[14] Q. Wu et al., “Deep Learning Methods for Predicting Disease Status Using Genomic Data.,” J. Biom. Biostat., vol. 9, no. 5, 2018.

[15] K. Raza, “Application Of Data Mining in Bioinformatics,” Indian J. Comput. Sci. Eng., vol. 1, no. 2, pp. 114–118, 2012.

[16] N. K. Sakthivel, N. P. Gopalan, and S. Subasree, “A Comparative Study and Analysis of DNA Sequence Classifiers for

Predicting Human Diseases," in Proceedings of the International Conference on Informatics and Analytics - ICIA-16, 2016, pp. 1–5.

[17] S. Sindhu and D. Sindhu, "International Journal of Computer Science and Mobile Computing Data Mining and Gene Expression Analysis in Bioinformatics," Int. J. Comput. Sci. Mob. Comput., vol. 6, no. 5, pp. 72–83, 2017.

[18] J. T. Wassan, H. Wang, and H. Zheng, "Machine Learning in Bioinformatics," in Encyclopedia of Bioinformatics and Computational Biology, Elsevier, 2018, pp. 300–308.

[19] Y.-R. Cho and M. Kang, "Interpretable machine learning in bioinformatics," Methods, vol. 179, pp. 1–2, Jul. 2020.

[20] C. M. Bishop, Pattern recognition and machine learning. springer, 2006.

[21] D. Medini et al., "Microbiology in the post-genomic era," Nat. Rev. Microbiol., vol. 6, no. 6, pp. 419–430, Jun. 2008.

[22] D. R. Schrider and A. D. Kern, "Supervised Machine Learning for Population Genetics: A New Paradigm," Trends Genet., vol. 34, no. 4, pp. 301–312, Apr. 2018.

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[24] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," Nat. Rev. Genet., vol. 16, no. 6, pp. 321–332, Jun. 2015.

[25] R. Guan et al., "Multi-label deep learning for gene function annotation in cancer pathways," Sci. Rep., vol. 8, no. 1, pp. 1–9, 2018.

[26] J. D. Wren, "A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide," Bioinformatics, vol. 25, no. 13, pp. 1694–1701, Jul. 2009.

[27] R. Upstill-Goddard, D. Eccles, J. Fliege, and A. Collins, "Machine learning approaches for the discovery of gene-gene interactions in disease data," Brief. Bioinform., vol. 14, no. 2, pp. 251–260, Mar. 2013.

[28] S. Oh, J. Lee, M.-S. Kwon, B. Weir, K. Ha, and T. Park, "A novel method to identify high order gene-gene interactions in genome-wide association studies: gene-based MDR," in BMC bioinformatics, 2012, vol. 13, no. 9, pp. 1–9.

[29] S. Sheehan and Y. S. Song, "Deep learning for population genetic inference," PLoS Comput. Biol., vol. 12, no. 3, p. e1004845, 2016.

[30] D. Capper et al., "DNA methylation-based classification of central nervous system tumours," Nature, vol. 555, no. 7697, pp. 469–474, Mar. 2018.

[31] R. Myers et al., "MA10.09 Evaluation of the Clinical Utility of the PanCan, EU-NELSON and Lung-RADS Protocols for Management of Screen Detected Lung Nodules at Baseline," J. Thorac. Oncol., vol. 14, no. 10, pp. S288–S289, Oct. 2019.

[32] E. J. Smith et al., "Expressed Sequence Tags for the Chicken Genome from a Normalized, Ten-Day-Old White Leghorn Whole Embryo cDNA Library. 2. Comparative DNA Sequence Analysis of Guinea Fowl, Quail, and Turkey Genomes," Poult. Sci., vol. 80, no. 9, pp. 1263–1272, Sep. 2001.

[33] J. Liu et al., "Genome-wide identification and validation of new reference genes for transcript normalization in developmental and post-harvested fruits of Actinidia chinensis," Gene, vol. 645, pp. 1–6, Mar. 2018.

[34] M. Peker, A. Arslan, B. Sen, F. V. Celebi, and A. But, "A novel hybrid method for determining the depth of anesthesia level: Combining ReliefF feature selection and random forest algorithm (ReliefF+RF)," in 2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA), 2015, pp. 1–8.

[35] R. Touati, I. Messaoudi, A. E. Oueslati, and Z. Lachiri, "Distinguishing between intra-genomic helitron families using time-frequency features and random forest approaches," Biomed. Signal Process. Control, vol. 54, p. 101579, Sep. 2019.

[36] R. Hu, G. Pei, P. Jia, and Z. Zhao, "Decoding regulatory structures and features from epigenomics profiles: A Roadmap-ENCODE Variational Auto-Encoder (RE-VAE) model," Methods, Oct. 2019.

[37] H.-C. Yi, Z.-H. You, D.-S. Huang, X. Li, T.-H. Jiang, and L.-P. Li, "A Deep Learning Framework for Robust and Accurate Prediction of ncRNA-Protein Interactions Using Evolutionary Information," Mol. Ther. - Nucleic Acids, vol. 11, pp. 337–344, Jun. 2018.

[38] C. Wang and J. Guo, "A data-driven framework for learners' cognitive load detection using ECG-PPG physiological feature fusion and XGBoost classification," Procedia Comput. Sci., vol. 147, pp. 338–348, 2019.

[39] D. L. Streiner and G. R. Norman, "'Precision' and 'Accuracy': Two Terms That Are Neither," J. Clin. Epidemiol., vol. 59, no. 4, pp. 327–330, Apr. 2006.

[40] U. O. F. Z. E. A. Piecemeal and T. O. P. R. Adenoma, "XIII National Congress of Digestive Diseases, Italian Federation of Digestive Diseases – FIMAD Palermo, 29 September – 3 October 2007," Dig. Liver Dis., vol. 39, pp. S139–S343, Sep. 2007.

[41] Y. Zhu et al., "Complete genome sequence and genome-scale metabolic modelling of Acinetobacter baumannii type strain ATCC 19606," Int. J. Med. Microbiol., p. 151412, Feb. 2020.

[42] U. O. F. Z. E. A. Piecemeal and T. O. P. R. Adenoma, "XIII National Congress of Digestive Diseases, Italian Federation of Digestive Diseases – FIMAD Palermo, 29 September – 3 October 2007," Dig. Liver Dis., vol. 39, pp. S139–S343, Sep. 2007.

Dr. Abdul Rauf Baig

Professor at the Dept. of Information Systems, College of Computer & Information Sciences, Al-Imam Muhammad bin Saud Islamic University (IMSIU), Riyadh, Saudi Arabia. He received his PhD in CS from Univ. of Rennes-I (France, April 2000), with 27 years of post-graduate experience, including 15 years of post-PhD teaching experience in Pakistan and Saudi Arabia. Researcher with 2 monographs, 40 journal publications, and 60 conference papers. PhD supervisor, approved by Higher Education Commission of Pakistan. Successfully supervised 7 PhD theses, several MS theses, research surveys, research projects, and undergraduate projects. His research interest includes Evolutionary Computation, Computational Intelligence, Particle Swarm Optimization, Machine Learning, Artificial Intelligence, Artificial Neural Networks, Soft Computing, Pattern Recognition.

Ms. May Almutairi

Information System master's student at the College of Computer & Information Sciences, Al-Imam Muhammad bin Saud Islamic University (IMSIU), Riyadh, Saudi Arabia. She received her bachelor's degree in Information Technology from King Saud University in 2014. Her research interest includes Artificial Intelligence (AI), Machine Learning, Predictive Models.