International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Classification of DNA Microarray Gene Expression Leukaemia Data through ABC and CNN Method

Abdul Wahid[1], M. Tariq Banday[2]

Submitted:25/03/2023 **Revised:**27/05/2023 **Accepted:**12/06/2023

Abstract: Biomedical and health-care informatics research is increasingly using big data technology. At an unprecedented velocity and scale, large volumes of biological and clinical data have been created and gathered. DNA microarray classification has been widely used in biological and medical research to study gene expression patterns, identify disease biomarkers, classify cancer subtypes, predict treatment responses, and discover novel gene functions. Predictive analytics is becoming more popular for its applications in healthcare and has a lot of potential. While the performance concerns are still in a way and optimization approaches are used to address these concerns. In the proposed methodology we have adopted hybrid approach of optimization algorithm (Artificial Bee colony) for feature selection and deep learning (Convolutional Neural Network) method for classification. ABC method helps in obtaining best features which also improves the accuracy of classification. The accuracy and other performance characteristics of the proposed algorithm CNN are examined. To demonstrate the usefulness of the proposed model, it is compared to various algorithms such as decision tree, random forest, and KNN based on performance metrics and the proposed approach achieves 98% accuracy which is remarkable as compared to other approaches.

Keywords: *Gene Expression, DNA Micro-array, Feature Selection, Swarm Intelligence, ABC Algorithm*

1. Introduction

The healthcare system demands minute attention on the selection of the data as well as the training process of the data. The integrated modern information technology had made it possible to predict and diagnose disease a time ahead to avail apt medical treatment and increase survival rates. The technological advancements are reflected by Electronic Health Records (EMRs), Electronic Medicine (e-Medicine), Mobile Health (M-Health) systems, etc. Due to the increasing volume of data, it is very difficult to predict anything from data manually. Algorithms have been proposed in the early stage of the healthcare prediction architecture and have gained revisions time by time. Machine learning automation will not only reduce the computation effort of the human but more patients would be cured with less computation complexity. The prediction architecture in any manner involves two phases namely the training and the classification [1]. The training process involves feature extraction, optimization, and usage of learning algorithms to train the system. Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. Feature extraction is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set. When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features [2]. The extracted features through the feature extraction algorithm are not always in the best shape and figure. It means that training through such algorithms will result in a non-efficient classification rate. In such a scenario, the optimization algorithm plays a vital role in order to normalize the extracted features. Normalization can be applied to the extracted features as well. In this case, the normalization will be termed optimization. It is not necessary that the extracted features are of best elementary value and hence optimization is a common frame of processing these days [3].The optimization algorithms can be of two categories such as Natural Computing and Swarm Intelligence. The Natural Computing is the type of algorithm which is directly inspired by the nature. It remains unaffected from any kind of load change or data distribution change. The issue with natural computing is that it consumes a lot of time for any processing. It is nowhere suitable when a quick analysis is required [4].TheSwarm Intelligence is the type of computing algorithm which is inspired by the behaviour of the elements which exist in nature. In addition, it is a behaviour that is observed mostly in a group. For example, Ant Colony Optimization (ACO), Artificial Bee Colony optimization

[1] *Research Scholar, Department of Electronics & Inst. Technology.University of Kashmir, INDIA*
[2] *Professor, Department of Electronics & Inst. Technology,University of Kashmir, INDIA*
Corresponding Author: abdul_wahid@zoho.com

(ABC), Cuckoo Search, Firefly, Whale Optimization, etc. [5].

The selection ofclassification algorithms are based on the computation complexity of the data. Healthcare data may be of the following architecture and types.

a) *Gene Expression Data:*It is a process by which a gene's DNA sequence is converted into the structures and functions of a cell. Non-Protein genes are not translated into protein. Genetic information, chemically determined by DNA structure is transferred to daughter cells by DNA replication and expressed by Transcription followed by Translation. The series of events is called "Central Dogma" is found in all cells and proceeds in similar ways except in retroviruses which possess a reverse enzyme transcriptase that converts RNA into complementary DNA. Biological information flows from DNA to RNA and from there to proteins [6].
b) *Other Data:*The data expressions that are not evaluated by gene expressions are called other types of data in health care. This type of data may contain ECG records, EEG records, EMG records, medical images, clinical trial results, etc.[7].

In addition to the existing approaches of categorization, this paper contributes to the following architecture of healthcare analysis.

a) Development of a novel behaviour of Swarm Intelligence based Artificial Bee Colony (ABC) algorithm for the selection of the attribute set of the gene expression data.
b) A design of the fitness function adopting the queen bee architecture for the selected attributes validation.
c) Training and classification using binary class and multi-class classifiers.

In the proposed study a two-stage hybrid approach is used. The first stage includes Artificial Bee Colony (ABC) algorithm as a selection criterion to select features and through this effective feature selection method, it is possible to select the relevant features faster. It also reduces the feature vector dimensions or matrix. The second stage includes applying classification and in this study Convolutional neural network is used as a classifier. After applying CNN classification method, the accuracy of the classifier is determined and compared with the other classifier. The dataset which is used for classification is gene expression monitoring via DNA microarray. This dataset provides a general approach to identify new cancer classes and assigning tumours to the known classes. The clinical diagnostic approach is one of the interesting areas of classification which predicts and classifies disease types and subtypes. Prevalent examples include diagnosis of leukaemia into ALL or AML [8] brain tumour diagnosis [9] and lymphoma [10] and breast cancer diagnosis [11]. DNA Microarray is usually microscope slides that are represented in printed form with thousands of tiny spots in defined position and each spot consists of known DNA sequence or gene [8].

Feature selection is the process of selecting distinguishing features from a set of features and removing irrelevant features. The aim of feature selection is to obtain important features and reducing the dimensions of the data. The ultimate impact of feature selection is to reduce processing time and to increase classification accuracy. First of all, feature selection process begins with sequential forward approach that is simplest greedy approach. It starts with empty set of features and sequentially adds features to get the highest number of features. To get the best performing features it needs to evaluate the feature subset and it is possible through the optimization of objective function. To optimize the objective function, the ABC algorithm is used for aiming the best feature selection.

ABC algorithm contains of three different artificial groups of bees. These groups are namely; employee, onlooker and scout bees. The onlooker bee waits for food source selection decision and it becomes employed bee once it gets the food source. The onlooker bee and employed bee are equal in numbers. Once the employed bee consumes the food source it becomes the scout bee and its job to do a random search for discovering new food sources. Hence in the perception of feature selection following pseudocode can be utilized.

Proposed Algorithm: Artificial Bee Colony for Feature Selection

```
While (Generation < 1000) do
    For every employed bee goes to the food source
    if Fitness (Current Food Source) > Fitness (New Food Source) then
    Keep the new found food source
    end if
```

```
        Every Onlooker Bee then probabilistically selects a food source to search around for    better food source(s)
        if Fitness (Current Food Source) > Fitness (New Food Source) then Keep the new found food source
        end if
        Scout Bees will search for new food sources once a previous food source is depleted
end while
```

2. Literature Review

Artificial Intelligence (AI) helps a key role in healthcare informatics in terms of identifying, tracking, and preventing various diseases using big genomic and clinical data. The success of the Human Genomic Project was a major paradigm in bioinformatics research followed by the conjugation of several ML approaches for the prediction of future trends and patterns. Traditionally, ML algorithms are implemented for model development based on learning from the existing data. Healthcare sector is benefitted from both clinical as well as genomic data that continuously generate a large volume of data that needs constant processing and interpretation. This had led to the usage of Big Data Analytics for handling massive medical data resulting in a big shift in healthcare [12].

An AI-based predictive clinical application to ease the overwhelmed emergency departments was developed by [13]. The major application of these predictive automatic tools was to predict mortality at the triage level and help in the categorization of the patients admitted to the emergency department. The ML model was programmed using Python and XGBoost library with the scikit-learn wrapper. The class imbalance was addressed with XGBoost and statistical analysis was performed using Python. The evaluation was performed against 799522 patient visits to an emergency department. The predictive model resulted in 0.962 AUC and 0.923 for early and short-term mortality, respectively. Overall, the gradient boosting model proved to exhibit high predictive efficacy at the emergency department.

Soffer et al [14] had integrated AI algorithms for Big Data processing to design models to access patients' risk stratification. The research was aimed to develop a mortality prediction model using Big Data Analytics. An ML model was constructed using laboratory tests, patient characteristics, comorbidities, and emergency department management data. Based on this raw information, the model was programmed using Python, XGBoost library with scikit. The statistical analysis was performed using Python that demonstrated that the designed ML model achieved AUC of 0.924 with a sensitivity of 0.88 and specificity of 0.83 with a false positive rate of 1:5.9.

The classification of the genes depends upon the features that are contributing in the classification. A feature combination and selection algorithm that incorporates machine learning architecture over gene expression data, is presented A hybrid architecture of meta-heuristics is implemented using soft-computing [15]**.** Therefore, involvement of large dataset representing cancerous classes is recommended [16].

The SVM based classification was performed with 87 genes which were selected using information gain proved to achieve highest AML and ALL class detection accuracies. The study concluded that gene selection followed by computational intelligence techniques could offer better diagnosis followed by treatment of patients [17]. In the same year, another study was proposed that was based on swarm-based architecture to offer cancer classification including leukaemia based on microarray gene expression data. Experimentation involved ABC for gene selection representing a particular cancer type from larger microarray dataset. The simulation analysis had shown that cancer classification using ABC with SVM demonstrated highest classification accuracy of 93.05% including 14 genes to discriminate cancer subjects into leukaemia classes (ALL and AML). This research had demonstrated high classification accuracy while considering a smaller set of genes that may incorporate false positive results. The training and the classification architecture remains the same whether the data belongs to Leukaemia or any other disease.

Dwivedi et al [18]evaluated neural network framework against five other machine learning techniques namely, Naïve Bayes, K-nearest neighbour, logical regression and SVM, demonstrating its outperformance with least erroneous detection at the testing stage. However, any feature extraction or feature reduction technique was not involved that raises the computation cost of the proposed design.

More recently, importance of feature selection was demonstrated by Arif and Shah for cancer classification based on microarray data. They had involved six-feature statistical approaches for feature selection in combination to PCA to address the high dimensionality issues while analysing microarray data. This was followed by the classification techniques namely, SVM, LDA and k-nearest neighbour. The simulation analysis has shown that PCA based feature analysis followed by SVM classifier outperformed the other combinations used in the study with correct recognition rate of 96%. The work utilized a binary classifier at classification stage and could be extended with the involvement of multiclass classifiers to improve the overall performance [19].

Microarray data is mostly considered as structured data and can have few or numerous features The large number of features put a serious challenge for research scientists in the field of machine learning and because of the likelihood of "false positives", which can perhaps occur at the time of building the model for prediction or during the selection of relevant features [20].

Zhou et al [21] proposed Multi-Layer Perceptron (MLP) neural network with two optimized metaheuristic algorithms, namely Artificial Bee Colony (ABC) and Particle Swarm Optimization (PSO) for estimating the heating load (HL) and cooling load (CL) of the energy efficient residential buildings. Three important criteria that includes mean absolute error (MAE), root mean square error (RMSE) and coefficient of determination are utilized to measure the accuracy of the models. The result proves ABC and PSO algorithms help the MLP to perform more efficiently.

Doreswamy et al [22] designed a Binary Bat Algorithm (BBA) based Feedforward Neural Network (FNN) hybrid model which taps the benefits of BBA and performance of FNN to classify three standard breast cancer datasets into malignant and benign cases. In this work, BBA was utilized to produce a V-shaped hyperbolic tangent function for training the network and a fitness function was employed for error minimization. FNNBBA based classification yielded 92.61% accuracy rate for training data and 89.95% for testing data.

Moreover, various hybrid evolutionary algorithms are available in the literature whichimproves the classification accuracy. The goal of many evolutionary algorithms is to seek for best features subset by adopting bio-inspired approaches such as particle swarm optimization (PSO), Honey Bee, Firefly algorithms and these algorithms perform better and require the involvement of experts to attain the desired performance outcome.

Alomari et al. [23]proposed Minimum Redundancy Maximum Relevancy (MRMR) which is a hybrid gene selection of filter technique and Bat algorithm (BA) using Microarray dataset which is a wrapper technique . MRMR was used to discover the most important genes from the entire dataset of gene expression and BA was used to search for the most informative gene subset from the reduced set produced by MRMR that aided in detecting cancers. The support vector machine (SVM) with 10-fold cross-validation was used as classifier method which helped in evaluating the BA. The three different Microarray datasets were used, which includes Colon, Breast, and Ovarian cancer datasets on the performance benchmark. Also the proposed approach was capable of finding the small size of gene subset with high classification rate by using Genetic algorithm (GA) to perform the comparison with MRMR-BA.

Tabares-Soto et al [24] proposed algorithms in classical ML and DL to classify the tumours described in the 11_tumor database. The authors obtained the performance accuracies between 90.6% for LR and 94.43% for CNN using k-fold cross-validation. Also, we show how a tuning process may or may not significantly improve algorithms' accuracies.An accurate and efficient classification method based on gene expression DNA microarray data and ML/DL methods that facilitates tumour type prediction in a multi-cancer-type scenario was demonstrated in results.

Hira et al [25] proposed a Map Reduced based parallel feature selection and extreme learning for micro-array cancer data classification. In the first phase gene expression dataset were pre-processed by attribute-wise normalization and then by setting thresholds on the original data. In the second phase a wrapper model was used through Adaptive Whale Optimization Algorithm (AWOA) with Nelder–Mead algorithm (NMA) to accomplish the feature subset selection.

Vaiyapuri et al[26] developed a red fox optimizer with deep-learning-enabled microarray gene expression classification (RFODL-MGEC) model which aims to improve classification performance by selecting appropriate features. Among the RFODL-MGEC model uses red fox optimizer (RFO)-based feature selection approach for deriving an optimal subset of features and. RFODL-MGEC model uses bidirectional cascaded deep neural network (BCDNN) for data classification. The parameters involved in the BCDNN technique were tuned using the Chaos Game Optimization (CGO) algorithm. The authors observed that RFODL-MGEC model accomplished superior results for subtype classifications.

Deng et al [27] proposed a two-phased gene selection approach by combining extreme gradient boosting (XGBoost) and a multi-objective optimization genetic algorithm (XGBoost-MOGA) for cancer classification in microarray datasets. In the first phase, an ensemble-based feature selection approach was used through XGBoost and in the second phase, XGBoost-MOGA searches for an optimal gene subset based on the most relevant genes groups by using a multi-objective optimization genetic algorithm. On doing comparative analysis of XGBoost-MOGA with other feature selection methods using two well-known learning classifiers on 13 publicly available microarray gene expression datasets and the result demonstrated that XGBoost–MOGA yields significantly better results in terms of various evaluation criteria, such as accuracy, F-score, precision, and recall.

Rostami et al [28] proposed a social network analysis-based gene selection approach that has two main objectives of the relevance maximization and redundancy minimization of the selected genes. Firstly a maximum community is selected repetitively and then among the existing genes an appropriate set of genes were selected by using the node centrality-based criterion. The authors reported that the results indicate better classification accuracy of microarray data and also decrease in time complexity.

Xie et al [29] proposed a new hybrid feature selection method Multi-Fitness RankAggreg Genetic Algorithm(MFRAG) which integrates nine feature selection methods. MFRAG works on the basis of the principle of “survival of the fittest”. The model improves the selection and mutation processes in genetic algorithms by enhancing the stability and reliability of the selection process through fusion mechanisms and integrated models It also guides the evolutionary process through a set of lists generated by a feature fusion model. Through experimental work and on doing the state-of-the-art analysis the MFRAG outperforms all standard methods in terms of classification accuracy and other evaluation criterion.

The literature survey had shown that recent time has witnessed a great shift in the healthcare domain due to technological advances. Further, it has been observed that AI approaches play a vital role in designing various automated predictive models to the power healthcare sector in terms of disease prediction and analysis.

3. Proposed Approach & Methodology

Based on the complexity of the evaluated features, machine learning algorithms are used for training and classification purposes. Due to its multi-label training architecture, Convolution Neural Network (CNN) has been observed as a key classification architecture that has been keenly adopted by researchers around the world [30] [31].

CNN is class of Artificial Neural Network (ANN) and systems are based on interconnected nodes or units that are also known as artificial neurons that are roughly equivalent to the neuron of the biological nervous system. The signals are processed through the neurons' connections that are equivalent to the synapses found in the brain. The layout of the artificial neural network is shown in Figure 1. A typical convolution neural network consists of the convolution layer, pooling layer, fully connected layers, and output with connections [32].

The architecture of this network has a five-layer input layer, a pooling layer (it can be more than one), and an output layer. It is also known as the Multilayer perceptron (MLP) due to the multiple numbers of pooling layers as shown in Figure 2.

The pooling layer can be called the distillation layer which distils some of the most important patterns from the inputs and forwards them to the next layer. Because of this layered network becomes faster and efficient through identifying the important data only.

The activation function has two important tasks:

a. It captures a non-linear relationship among the inputs.
b. It transforms the input into a more helpful output.

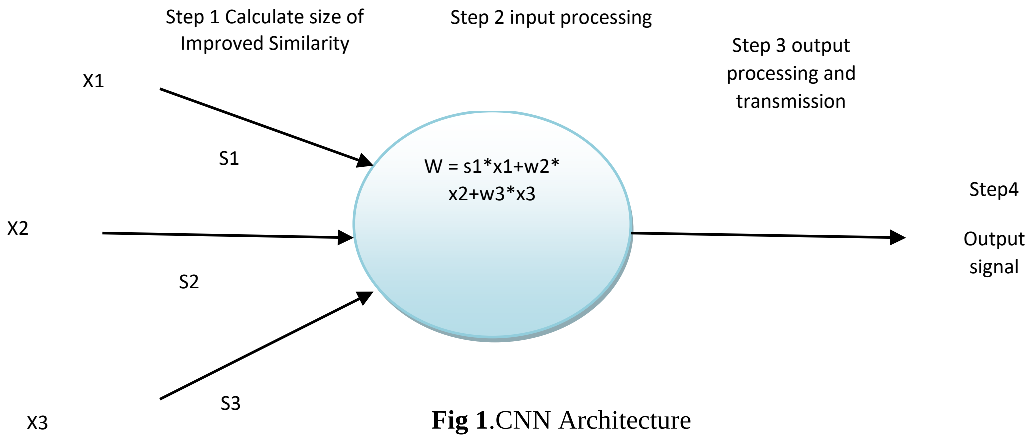
Step 1 Calculate size of Improved Similarity
Step 2 input processing
Step 3 output processing and transmission
X1
S1
X2
S2
X3
S3
W = s1*x1+w2* x2+w3*x3
Step4
Output signal

Fig 1.CNN Architecture

The activation function requires the propagation model which is also termed as sigmoid function. The sigmoid can be linear or polynomial based on the linearity defined in the data. Sigmoid activation function generates an output between 0 and 1. The pooling layer provides the last prediction through the output layer. One of the specialties of convolution networks is that the hidden unit's factors. The network which implemented a convolution network in it can extract higher-order statistics by adding one or more pooling layers. As the training process requires the ground truth in order to train the system, clustering becomes vital if the ground truth is not provided along with the data.

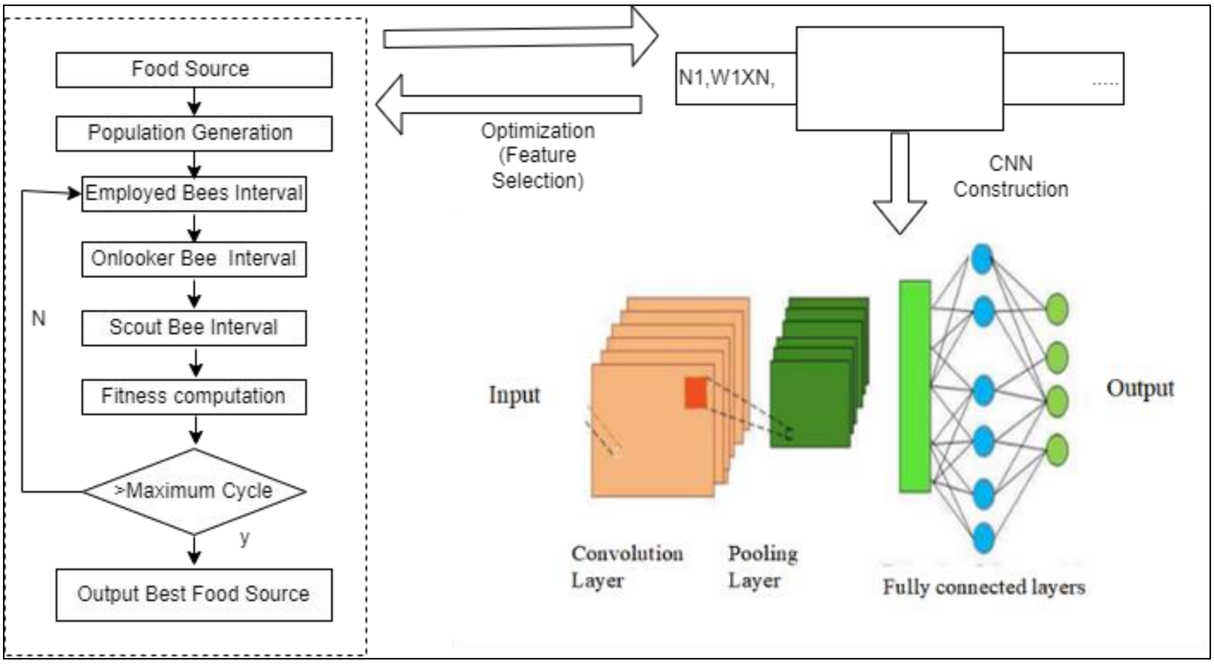

Fig 2: Proposed Implementation Architecture

A novel hybrid algorithm has been proposed using the ABC and classification using the CNN technique. The proposed technique is based on generating the reward using the employee bee and onlooker bee. The attribute set has been extracted using the employee bee and reward points are generated using the total levy flight work by the bees. The maximum fitness is obtained when the maximum efficiency shown by the bees and thus awarded by the reward. The ABC algorithm utilized a reward mechanism as illustrated earlier in the segment. The reward mechanism recursively tests the acceptance and rejection of the data element in the provided category set. The fitness function has also been evaluated multiple times to set the threshold of the processing through ABC algorithm. The fitness function is dynamic in nature and adapts a new threshold every time it is supplied with some set of input data. The functionality modifies the data as per its ground truth value and due to being generic in nature, it also supports 'n' number of classes to be processed. Further the selected features are provided to classify the data to CNN method.

In order to validate the proposed work, we have used quantitative parameters like classification accuracy, precision, recall, etc.[33][34]. The classification architectures do not have much of a possibility to change the processing elements and hence optimality in the training can only be attained prior to passing the data to the process engine. This research article majorly focuses on the enhancement of the feature selection and attribute set selection to enhance the overall classification accuracy. The implementation architecture is developed in python (figure 2). The designing has been done to run applications that are of high performance.

Artificial Bee Colony Algorithm utilizes for feature selection. As the training architectures for Deep Neural Networks (DNNs) do not provide much flexibility other than providing variations in the propagation layer, the only way to optimize the training is to provide it the best segmented data.

4. Results and Discussion

The proposed algorithm has been developed using the swarm technique ABC and CNN. The performance metric has been evaluated by measuring the prediction accuracy, true positive rate (TPR) and false positive rate (FPR).

- Classification Accuracy

It is the most common metric used to estimate the efficiency and performance of the classification algorithm. The classification accuracy is defined in the given equation.

$$Accuracy = \frac{T_{negative} + T_{positive}}{T_{negative} + F_{negative} + T_{positive} + F_{positive}}$$

- TPR (Recall)

It is the ration of true positive rate to the sum of true positive and false negative. It is also called as recall and sensitivity. Its value lies in the range of 0 – 100. The closer value to 100 signifies the better positive rate.

$$TPR\ (Recall) = \frac{T_{positive}}{T_{Positive} + F_{negative}}$$

- FPR

It is also called as False Alarm Rate. It is defined as the ratio of false positive to the sum of false positive and true negative. In other words, it is the probability to reject the falsifying events.

$$FPR = \frac{F_{positive}}{F_{positive} + T_{negative}}$$

Table 1. Estimation of Accuracy using the optimization and without optimization techniques

Simulation Rounds	Accuracy with ANN (%)	Accuracy with CNN (%)	Accuracy with CNN and ABC (%)
100	84.32	91.5435	96.6453
200	84.3232	91.743932	97.09432
300	84.42644	91.944365	97.44334
400	84.52969	92.144797	97.79236
500	84.63293	92.34523	98.14138
600	84.73617	92.645662	98.5904
700	84.83942	92.946094	98.93942
800	84.94266	93.246527	99.28844
900	85.0459	93.546959	99.43746
1000	85.14914	93.847392	99.58648

Table 1 shows the accuracy using the optimization technique and without optimization approach. The analysis results showed that leukaemia diseases predicted in the healthcare big data. The accuracy of the predicted diseases has been tested using the CNN and ABC optimization approach. The results showed that average accuracy with ANN is about 85% while that of other technique such as using CNN only, it is 93.22% and using CNN and ABC, it is about 98.3%. Thus, proposed system shows effective results to predict the diseases in healthcare Big Data. The graphical representation in terms of average accuracy is shown in Fig. 3.

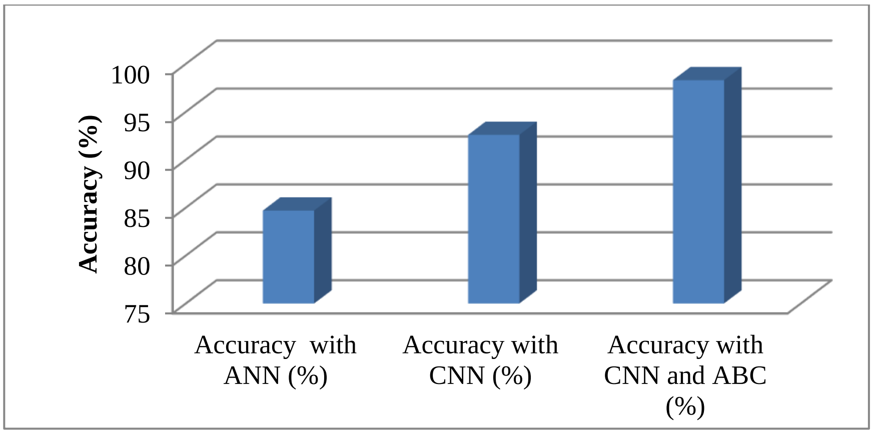

Fig. 3 Comparison of Accuracy using the proposed technique

Fig. 3 shows the average accuracy using the ANN and with CNN and ABC. The outcomes showed that average accuracy using the CNN and optimization techniques provide better results in comparison to without using CNN and optimization approach.

Table 2. Estimation of TPR using the optimization and without optimization techniques

Simulation Rounds	TPR with ANN (%)	TPR with CNN (%)	TPR with CNN and ABC (%)
100	86.32	90.5435	95.9453
200	86.3232	90.74393	96.39432
300	86.42644	90.94436	96.74334
400	86.52969	91.1448	97.09236
500	86.63293	91.34523	97.44138
600	86.73617	91.64566	97.8904
700	86.83942	91.94609	98.23942
800	86.94266	92.24653	98.58844
900	87.0459	92.54696	98.73746
1000	87.14914	92.84739	98.88648

Table 2 shows the TPR using the ANN and without optimization approaches. The analysis results showed that leukaemia diseases predicted in the healthcare big data. The True positive rate of the predicted diseases has been tested using the CNN and ABC optimization approach. The results showed that average TPR with using ANN is about 85% while that of other technique such as using CNN only is 93.22% and using CNN and ABC is about 98.3%. Thus, proposed system shows effective results to predict the diseases in healthcare Big Data. The graphical representation in terms of average TPR is shown in Fig. 4.

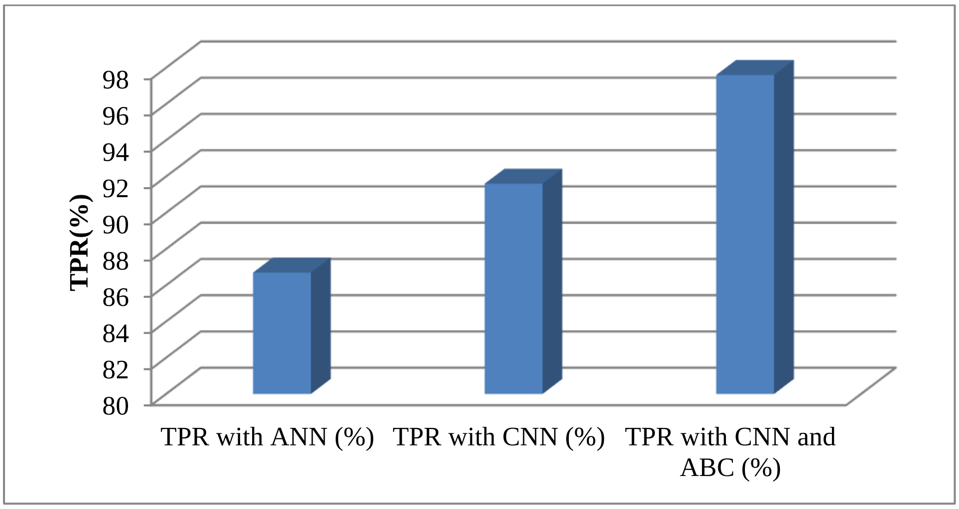

Fig. 4 Average TPR without using the optimization and with CNN and ABC

Table 3. Estimation of Recall using the optimization and without optimization techniques

Simulation Rounds	Recall with ANN (%)	Recall with CNN (%)	Recall with CNN and ABC (%)
100	89.32	91.5435	96.9453
200	89.3232	91.74393	97.19432

300	89.42644	91.94436	97.54334
400	89.52969	92.1448	97.89236
500	89.63293	92.34523	98.24138
600	89.73617	92.64566	98.6904
700	89.83942	92.94609	98.83942
800	89.94266	93.24653	98.98844
900	90.0459	93.54696	99.13746
1000	90.14914	93.84739	99.28648

Table 3 shows the computation of Recall using the optimization technique and without optimization approaches. The analysis results showed that leukaemia diseases predicted in the healthcare big data. The recall value of the predicted diseases has been tested using the CNN and ABC optimization approach. The results showed that average recall with ANN is about 89.6% while that of other technique such as using CNN only, it is 92.59 % and using CNN and ABC is about 98.27%. Thus, proposed system shows effective results to predict the diseases in healthcare Big Data.

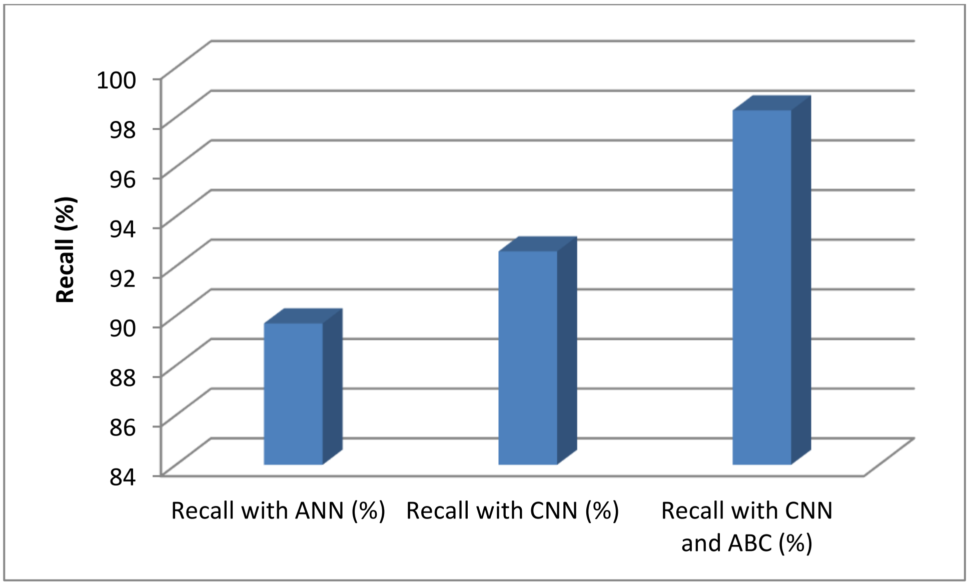

Fig. 5 Average Recall without using the optimization and with CNN and ABC

Fig. 5 shows the average recall using the optimization technique and with CNN and ABC for ALL and AML leukaemia disease. The outcomes showed that average recall using the CNN and optimization techniques provide better results in comparison to with ANN and optimization approach.

Table 4. Estimation of FPR using the optimization and without optimization techniques

Simulation Rounds	**FPR with ANN (%)**	**FPR with CNN (%)**	**FPR with CNN and ABC (%)**
100	0.7584	0.54576	0.38675
200	0.75609	0.50233	0.36554
300	0.75378	0.4589	0.34433
400	0.75147	0.43547	0.32312
500	0.74916	0.41204	0.30191
600	0.74685	0.40861	0.2807

700	0.74454	0.40518	0.25949
800	0.74223	0.40175	0.23828
900	0.73992	0.39832	0.21707
1000	0.73761	0.39489	0.19586

Table 4 shows the FPR using the optimization technique and without optimization approaches. The analysis results showed that leukaemia diseases predicted in the healthcare big data. The False positive rate of the predicted diseases has been tested using the CNN and ABC optimization approach. The results showed that average FPR with ANN is about 0.74 while that of other technique such as using CNN only, it is 0.43 and using CNN and ABC, it is about 0.29. Thus, proposed system shows effective results to predict the diseases in healthcare Big Data. The graphical representation in terms of average FPR for leukaemia disease is shown in Fig. 6.

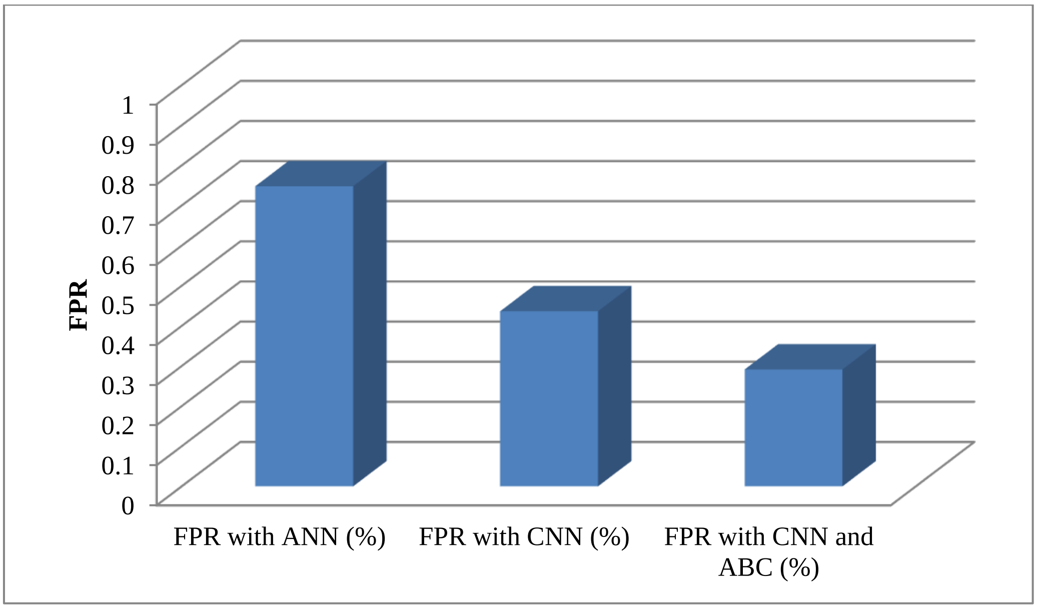

Fig. 6 Average FPR without using the optimization and with CNN and ABC

Comparative Analysis

This section compares the proposed approach with the existing techniques.

Table 5. Comparison of the proposed and existing technique

Parameters	Proposed technique (%)	Existing technique (%)	
	CNN+ABC	Dwivedi, A. K. (2018) (Artificial Neural Network)	Ayyad et al., (2019) (5NN)
TPR	98.88	100	70.2
Classification Accuracy	99.58	98	70

The above table shows the comparison of the proposed technique and existing technique to determine the TPR (Recall) value and the classification (predictive) accuracy. The outcomes show that TPR has been improved by 28.68% in comparison to[35]while that of other existing approach [18]it is less than the existing method. Consequently, the Classification (prediction) accuracy of the proposed technique has been improved by 1.58% from [18] and29.58% from[35].

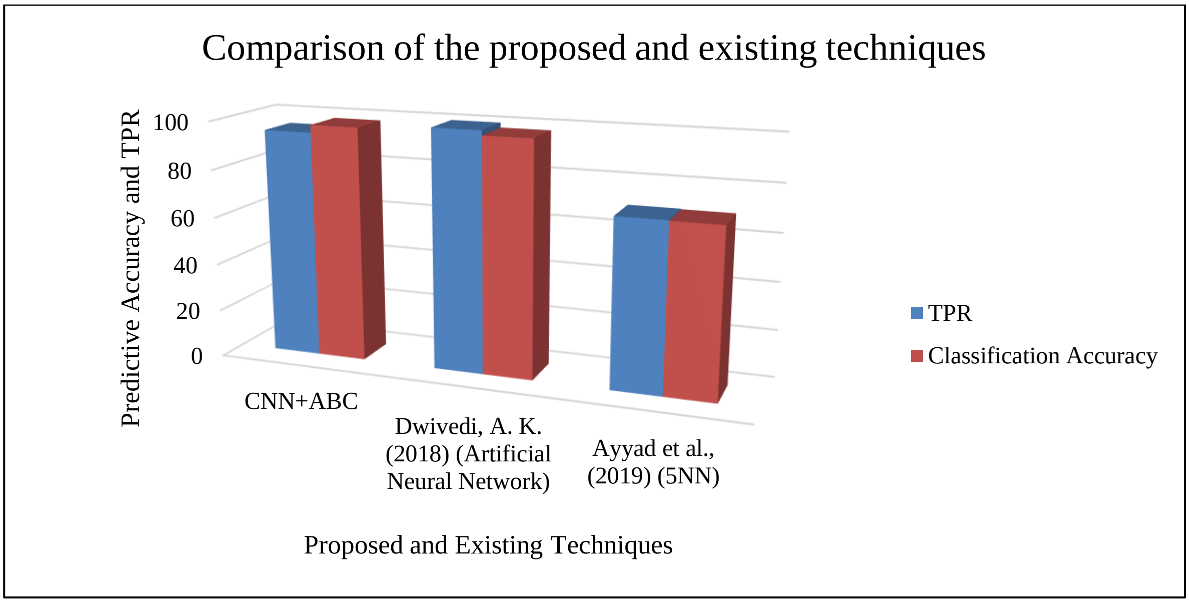

Fig 7 Comparison of the proposed and existing techniques to determine the predictive accuracy& TPR

The given figure shows that the proposed approach shows better results in contrast to the existing techniques. The results show that the proposed technique has predictive accuracy 99.58% while that of existing techniques, it is 98% and 70%. Thus, accuracy has been improved by applying the proposed hybrid approach.

5. Conclusion

The major goal of this research is to use a prediction model in healthcare. In this study, we employed a Swarm Intelligence-based Artificial Bee Colony (ABC) algorithm which involves the creation of a fitness function for the validation of the required attributes based on the queen bee architecture.The study focused on Leukaemia types ALL and AML from gene Expression DNA Micro-array dataset. First ANN and then CNN's accuracy was demonstrated and the CNN with ABC was performed and which gave better results in comparison to individual approaches. On the basis of performance indicators, the suggested model is compared to several other algorithms such as decision tree, random forest, and KNN to illustrate its use. As a result, the suggested approach achieves98 percent accuracy and when compared to other algorithms, the proposedmethodproves to bebetter and effective in predicting the illnesses of interest.

Competing Interests

First Author receives the grant from under FDP scheme of UGC India.

Second Author has no conflict of Interest.

Author Contribution Statement

Study conception and design: First Author, Second Author;

Data collection: First author;

Analysis and interpretation of results: First Author;

Draft manuscript preparation: Second Author.

Ethical and informed Consent for Data Used

All data were collected and handled in accordance with ethical standards, including anonymization and secure storage, to ensure the protection of participants' privacy and confidentiality.

Funding

UGC India under Faculty Development Programme for First Author

Not Applicable for Second Author

Data Availability and Access

The dataset used and analyzed during the current study is Gene expression dataset (Golub et al.)

Molecular Classification of Cancer by Gene Expression Monitoring and is available athttp://www-genome.wi.mit.edu/MPR/ .The data is available publicly and can be used by the machine learning community for the empirical analysis of machine learning algorithms.

References

[1] Gambhir, S., Malik, S. K., & Kumar, Y. (2016). Role of soft computing approaches in healthcare domain: a mini review. *Journal of medical systems*, *40*(12), 1-20.

[2] Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., ... & Lee, S. I. (2018). Explainable machine-learning predictions for the

prevention of hypoxaemia during surgery. *Nature biomedical engineering*, *2*(10), 749-760.

[3] Khan, S., Khan, A., Maqsood, M., Aadil, F., & Ghazanfar, M. A. (2019). Optimized gabor feature extraction for mass classification using cuckoo search for big data e-healthcare. *Journal of Grid Computing*, *17*(2), 239-254.

[4] Chawda, B., & Patel, J. (2016). Natural Computing Algorithms–A Survey. *International Journal of Emerging Technology and Advanced Engineering*, *6*(6).

[5] Nayar, N., Ahuja, S., & Jain, S. (2019). Swarm intelligence for feature selection: a review of literature and reflection on future challenges. *Advances in data and information sciences*, 211-221.

[6] Wang, L., Xi, Y., Sung, S., & Qiao, H. (2018). RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes. *BMC genomics*, *19*(1), 1-13.

[7] Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, *6*(1), 1-25.

[8] Golub, T. R., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science (80-.), 286(5439), 531–537.

[9] Wang, J., Bø, T. H., Jonassen, I., Myklebost, O., & Hovig, E. (2003). Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. BMC Bioinformatics, 4(1), 60

[10] Kim, J., et al. (2005). Identification of potential biomarkers of genotoxicity and carcinogenicity in L5178Y mouse lymphoma cells by cDNA microarray analysis. Environmental and Molecular Mutagenesis, 45(1), 80–89

[11] Hambali, M., Saheed, Y., Oladele, T., & Gbolagade, M. (2019). ADABOOST ensemble algorithms for breast cancer classification. Journal of Advance Computer Research, 10(2), 31–52.

[12] Hassan, M. K., El Desouky, A. I., Elghamrawy, S. M., & Sarhan, A. M. (2019). Big data challenges and opportunities in healthcare informatics and smart hospitals. In *Security in smart cities: Models, applications, and challenges* (pp. 3-26). Springer, Cham.

[13] Klug, M., Barash, Y., Bechler, S., Resheff, Y. S., Tron, T., Ironi, A., ... & Klang, E. (2020). A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a nine-point triage score. *Journal of general internal medicine*, *35*(1), 220-227.

[14] Soffer, S., Klang, E., Barash, Y., Grossman, E., & Zimlichman, E. (2021). Predicting In-Hospital Mortality at Admission to the Medical Ward: A Big-Data Machine Learning Model. *The American journal of medicine*, *134*(2), 227-234.

[15] Alonso-González, C.J., Moro-Sancho, Q.I., Simon-Hurtado, A. and Varela-Arrabal, R., 2012. Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods. *Expert Systems with Applications*, *39*(8), pp.7270-7280.

[16] Alshamlan, Hala M., Ghada H. Badr, and Yousef A. Alohali. "Abc-svm: artificial bee colony and svm method for microarray gene selection and multi class cancer classification." *Int. J. Mach. Learn. Comput* 6, no. 3 (2016): 184.

[17] Sheikhpour, R., and M. Aghaseram. "Diagnosis of acute myeloid and lymphoblastic leukemia using gene selection of microarray data and data mining algorithm." *Scientific Journal of Iran Blood Transfus Organ* 12, no. 4 (2016): 347-357.

[18] Dwivedi, Ashok Kumar. "Artificial neural network model for effective cancer classification using microarray gene expression data." *Neural Computing and Applications* 29, no. 12 (2018): 1545-1554.

[19] Arif, Muhammad Azharuddin, and Zuraini Ali Shah. "Implementation of Statistical Feature Selection and Feature Extraction on Cancer Classification." *Academia of Intelligence Computing* 1, no. 1 (2020): 21-29.

[20] Remeseiro López, B., & Bolon Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in Biology and Medicine, 112

[21] Ashok Kumar, L. ., Jebarani, M. R. E. ., & Gokula Krishnan, V. . (2023). Optimized Deep Belief Neural Network for Semantic Change Detection in Multi-Temporal Image. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2), 86–93. https://doi.org/10.17762/ijritcc.v11i2.6132

[22] Guofeng Zhou et al, Employing artificial bee colony and particle swarm techniques for optimizing a neural network in prediction of heating and cooling loads of residential buildings, Journal of Cleaner ProductionVolume 254, 1 May 2020, 120082.

[23] Doreswamy, & UmmeSalma, M. (2016). A binary bat inspired algorithm for the classification of breast cancer data. International Journal on Soft

Computing Intelligence and Applications, 5(2/3), 1–21.

[24] Alomari, O. A., et al. (2017). Mrmr ba: A hybrid gene selection algorithm for cancer classification. *Journal of Theoretical and Applied Information Technology, 95*(12), 15

[25] Tabares-Soto, R.; Orozco-Arias, S.; Romero-Cano, V.; Segovia Bucheli, V.; Rodríguez-Sotelo, J.L.; Jiménez-Varón, C.F. A Comparative Study of Machine Learning and Deep Learning Algorithms to Classify Cancer Types Based on Microarray Gene Expression Data. PeerJ Comput. Sci. 2020, 6, e270.

[26] Hira, S.; Bai, A. A Novel Map Reduced Based Parallel Feature Selection and Extreme Learning for Micro Array Cancer Data Classification. Wirel. Pers. Commun. 2022, 123, 1483–1505.

[27] Vaiyapuri, T.; Liyakathunisa; Alaskar, H.; Aljohani, E.; Shridevi, S.; Hussain, A. Red Fox Optimizer with Data-Science-Enabled Microarray Gene Expression Classification Model. Appl. Sci. 2022, 12, 4172.

[28] Deng, X.; Li, M.; Deng, S.; Wang, L. Hybrid Gene Selection Approach Using XGBoost and Multi-Objective Genetic Algorithm for Cancer Classification. Med. Biol. Eng. Comput. 2022, 60, 663–681.

[29] Rostami, M.; Forouzandeh, S.; Berahmand, K.; Soltani, M.; Shahsavari, M.; Oussalah, M. Gene Selection for Microarray Data Classification via Multi-Objective Graph Theoretic-Based Method. Artif. Intell. Med. 2022, 123, 102228.

[30] Xie, W.; Fang, Y.; Yu, K.; Min, X.; Li, W. MFRAG: Multi-Fitness RankAggreg Genetic Algorithm for Biomarker Selection from Microarray Data. Chemom. Intell. Lab. Syst. 2022, 226, 104573.

[31] Zellar, P. I. . (2021). Business Security Design Improvement Using Digitization. International Journal of New Practices in Management and Engineering, 10(01), 19–21. https://doi.org/10.17762/ijnpme.v10i01.98

[32] Ngiam, K. Y., & Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, *20*(5), e262-e273.

[33] Yu, H., Yang, L. T., Zhang, Q., Armstrong, D., & Deen, M. J. (2021). Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. *Neurocomputing*, *444*, 92-110.

[34] Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938.

[35] Talasila, V., Madhubabu, K., Madhubabu, K., Mahadasyam, M., Atchala, N., & Kande, L. (2020). The prediction of diseases using rough set theory with recurrent neural network in big data analytics. *International Journal of Intelligent Engineering and Systems*, *13*(5), 10-18.

[36] Jayasri, N. P., & Aruna, R. (2021). Big data analytics in health care by data mining and classification techniques. *ICT Express*.

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