Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777725

GENE EXPRESSION DATA CLASSIFICATION USING ARTIFICIAL INTELLIGENCE AND DATA MINING TECHNIQUE

Sivasubramaniam S
Research Scholar, PG and Research Department, Department of Computer Science
Park's College, Chinnakarai, Tirupur, India.

Dr. S. Prabhu
Assistant Professor, Department of Computer Science, Government Arts & Science College
Thittamalai, Nambiyur, India.

Abstract

AI (Artificial intelligence) developments in recent years have converted the traditional healthcare system into smart healthcare. Medical services may be enhanced by embracing important technology like AI. The healthcare sector has a variety of opportunities because to AI convergence. Moreover, data mining has been crucial in revealing hidden patterns in huge datasets. The current method has been shown to have issues with reduced classification accuracy in earlier studies. In this study, AFOECNN (Adaptive Firefly Optimization with Enhanced Convolution Neural Networks) is developed to increase the classification accuracy in order to address the aforementioned issue. Pre-processing, choosing feature subsets, and classification are the three primary stages of the proposed approach. KMC (K-Means Clustering) technique is used for pre-processing in order to reduce the noise data from the provided gene expression dataset. By employing k-means centroid values to manage missing features, it more effectively raises classification accuracy. The pre-processed features are employed in the feature subset selection approach to extract more beneficial features from the cancer dataset. Based on the best fitness values, the vital and obvious characteristic is computed using the objective function. It is carried out using the AFO (Adaptive Firefly Optimisation) algorithm. The data is transformed into pictures, and then augmentation is applied to those images. Finally, using a training and testing model, the ECNN algorithm is employed for classification. Using weight values, it more correctly categorises the tumour traits. This work's experimental ioutcomes show that AFOECNN is better than existing algorithms in terms of precision, recall, f-measure, accuracy values and reduced time complexities.

Key words: Gene expression data, Artificial Intelligence (AI), data mining, Adaptive Firefly Optimization with Enhanced Convolution Neural Network (AFOECNN) algorithm

1. Introduction

Microarray data classifications are supervised learning tasks that diagnose sample's categories expression array phenotypes [1]. They create a classifier model from labelled gene expression data samples to classify fresh data samples into various illnesses. The simultaneous monitoring of thousands of genes' levels of expression throughout critical biological processes and across groups of related samples is possible with DNA microarray technology. As microarray data analysis is beneficial for classifying illnesses according to their phenotypes, its knowledge is becoming more and more significant.

The biggest problem today is how to successfully incorporate genetics into precision medicine that applies to people of all ancestries, with a variety of illnesses, and in other distinct groups. To do this, it will be necessary to make intelligent use of AI and MLTs (machine learning techniques). AI-based techniques have become effective tools to revolutionise healthcare [2]. In recent years, the healthcare industry began utilising information technology to create cutting-edge apps and improve the diagnostic and therapeutic procedures. The main forces behind the production of enormous amounts of digital data are cutting-edge methods and scientific theories. Advanced clinical applications are the products of information technology that has recently been produced after that.

Physicians typically employ hypothetico deductive reasoning to develop a diagnosis for every specific patient. The doctor begins by addressing the primary problem before posing pertinent questions about it. The doctor creates a differential diagnosis from this initial, limited feature set and then chooses which features (historical inquiries, physical exam results, laboratory tests, and/or imaging investigations) to gather next in order to confirm or exclude the diagnoses in the differential diagnostic set. The most helpful elements are chosen, and the procedure is ended and the diagnosis is approved when the likelihood of one of the diagnoses exceeds a specified threshold of acceptance. With just a few traits, it could be feasible to make a diagnosis with an acceptable degree of certainty without needing to process complete feature sets.

Data mining is currently a crucial computational strategy for applications in the field of medicine. Examples of the development of data mining applications and their effects include managing healthcare data, epidemiologies, taking care of critical/non-critical patients, extracting information using image analysis and automatic identifications of of diseases by healthcare administrations. Useful information and concealed knowledge have been discovered from medical test data [3] [4] by data mining methods. Cancer researches employ a number of data mining techniques including classifications, clustering, predictions, association rules, decision trees, and neural networks.

Finding effective feature subsets (discriminated features), as well as enhancing dataset quality for better and quicker outcomes, need feature selection. To provide improved representations, several subsets of the retrieved characteristics from the provided datasets have been assessed. The effectiveness of breast cancer classifiers depends on both feature selection techniques and classification algorithms [5]. The classifier may become confused and produce inaccurate results for the supplied breast cancer dataset if improper and irrelevant characteristics are used. Feature subset selection, which eliminates the unnecessary and duplicate characteristics from the original breast cancer dataset, is the optimization-based solution to this issue [6]. While estimating kernel density for the purpose of diagnosing breast cancer, it chooses the feature subset and sets the kernel bandwidth.

The main focus of the research is on categorising gene expression data using AI and data mining techniques. Although numerous studies and methodologies have been created, the accuracy of the categorization of gene expression has not significantly increased. The disadvantages of the current methodologies include the time commitment and inaccurate cancer dataset categorization results. In order to overcome the aforementioned issues, this research

recommends adopting AFOECNN for enhanced performances. Important contributions of this study include pre-processes of data, selecting feature subsets using AFO and classifications using ECNN. The suggested method employs effective algorithms with the provided gene expression dataset to produce more accurate results.

The rest of this paper is structured as follows: literature review of gene expression datasets are described in Section 2. The suggested approach for classifying gene expression datasets using AI and data mining is presented as Section 3 followed by experimental findings in Section 4. The paper is concluded in Section 5.

2. Related work

Singh and Shailendra (2017) [7] presented in their study an innovative algorithm where optimal genes counts were obtained by the usage of different data reduction filters where low variances and low entropies were applied resulting in accurate inputs for further gene selection methods. They tested their method on identifications of prostate, breast, and lung cancers from gene expression datasets for. They executed gene deduction analyses to determine maximal identifications of specific groups of cancer genes.

To choose a subset of features, Chowdhury et al. (2019) suggested novel feature selections based on RNNs (recurrent neural networks). They used the technique specifically for selecting features from micro-array cell data. They built selection models within the framework (4) using various architectures of RNN including recurrent units, LSTM (long short-term memory), , and bi-directional LSTM. The study demonstrated system's efficacy real-world micro-array datasets.

Danaee et al. (2017) proposed in [9], DLTs (deep learning techniques) for detecting cancers by their identifications of breast cancer genes in diagnostics. The study discovered functional characteristics from high-dimensional gene expression patterns using SDAE (Stacked Denoising Autoencoders). The efficacy of their extracted representations were classified to confirm its utility in cancer diagnosis. Finally, by studying the SDAE connection matrices, we discovered a collection of highly interacting genes. The findings and analyses show that these highly interacting genes may serve as breast cancer indicators that need further investigation.

Sawhney et al (2018) In [10], investigated the use of Random Forest classifiers in diagnosing cancers of the breasts and livers, Cervical/Hepatocellular Carcinomas with enhanced classification accuracies and feature reductions when compared to other similar methods. Their addition of penalties to fitness functions promoting Binary FA (Firefly Algorithm), resulted in drastically reducing optimal subsets.

Zhang et al. (2017) in [11] studied the issue of feature subset selection using a novel self-adaptive FA.. Traditional FA traps fireflies in limited areas without giving them the chance to explore new search spaces since it relies on consistent control settings to solve various challenges. The study used two unique parameter selection algorithms that dynamically managed light absorption coefficients and controlled randomization parameter for handling limitations of FA. The goal function has a significant impact on the choice of features, which is another crucial aspect of the feature subset selection problem. In this study, we employ

mutual information-based criteria that not only measured correlations between two variables but also determined emendation of features amongst FA selected feature subsets. Using a number of benchmark datasets, the method is contrasted with DE (differential evolution), GA (genetic algorithms), and two iterations of PSO (particle swarm optimisation). The outcomes show how effective and competitive the strategy is in terms of classification accuracy and computing performance.

FCLF-CNN (completely-connected layer first Convolution Neural Networks) a novel CNN that embedded fully connected layers before first convolution layers was proposed by Liu et al., 2018) in [12]. The study used fully connected layers as approximators or encoders to transform raw data into localised representations. The study trained on 4 different FCLF-CNN types got their generation of FCLF-CNN ensembles. The study found improved performances while using 5-fold cross validations on the results of WDBC and WBCD datasets. FCLF-CNN outperformed pure MLPs (multi-layer perceptrons) and pure CNN on both datasets in terms of improved classifications.

To best categorise pulmonary nodules with CNNs, Pfeffer et al. (2022) adopted the evolutionary technique described in [13]. GA, which employs a variety of bio-inspired natural selection and Darwinian strategies to approach optimal solutions, is used for the development of CNN structures and hyperparameter optimisation. Deep Local-Global Network and FractalNet were two hand designed deep learning models that were trained for comparison. The results show that the GA-CNN with the best fit (91.3%) outperformed handcrafted models where their experimental findings suggested that usage of GAs would be useful for diagnostics, and that future complete automations using GAs may result from their designs and optimizations for a variety of clinical applications.

3. Proposed methodology

AFOECNN proposed in this work enhanced classifications gene expressions from datasets. The proposed system includes feature subset selection using the AFO algorithm, preprocessing using the KMC technique, and classification using the ECNN algorithm. Fig. 1 displays the block diagram in its entirety.

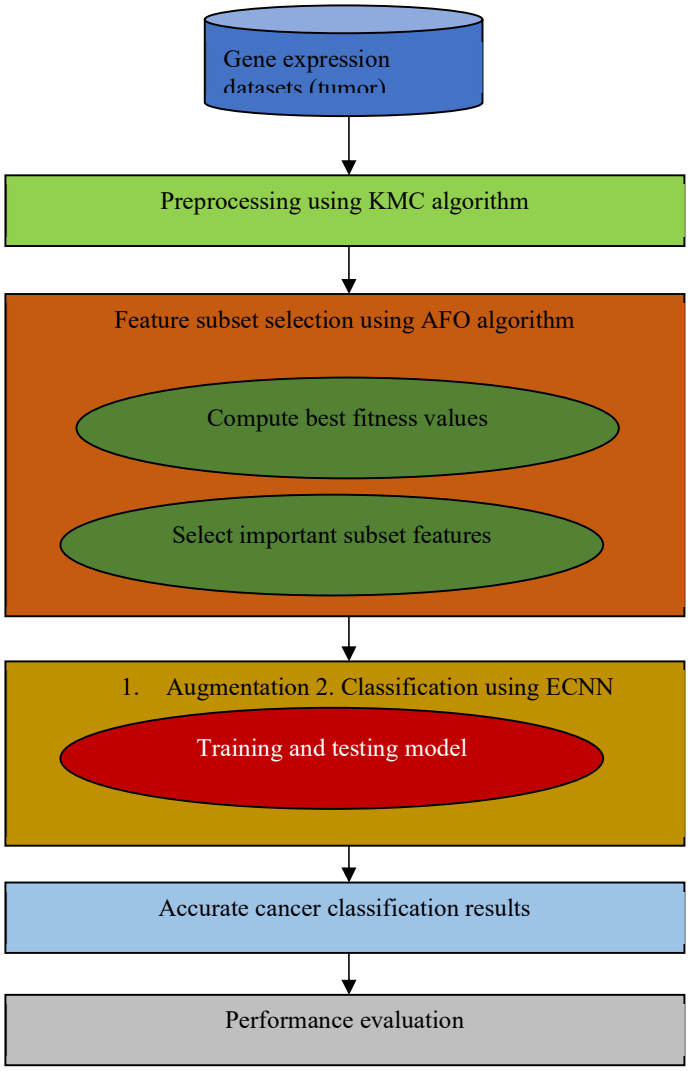

Fig 1 Overall block diagram of the proposed system

3.1 Pre-processing via KMC algorithm

The KMC approach is used in this study's pre-processes to increase gene expression dataset's classification precisions. It works very hard to complete missing values, lessen noise, and correct errors in the dataset. Because it is based on initial centroids of clusters, KMC is a useful clustering technique for classifying comparable data into groups [14]. Cluster centroids are determined using Euclidean distances. This approach starts with random partitions and computed cluster centres consistently or cluster vector average, and data items are reassigned to clusters whose cluster centres are nearest to them. When there are no more reassignments, the procedure is terminated. Intra-cluster variances or sums of squares of attribute and their corresponding cluster centre differences get locally minimized. The advantages of KMC include their simplicity in implementations and linear runtimes even when data component counts increase. This work maintains different clusters for different classes. A sample of the KMC algorithm is shown in Fig. 2.

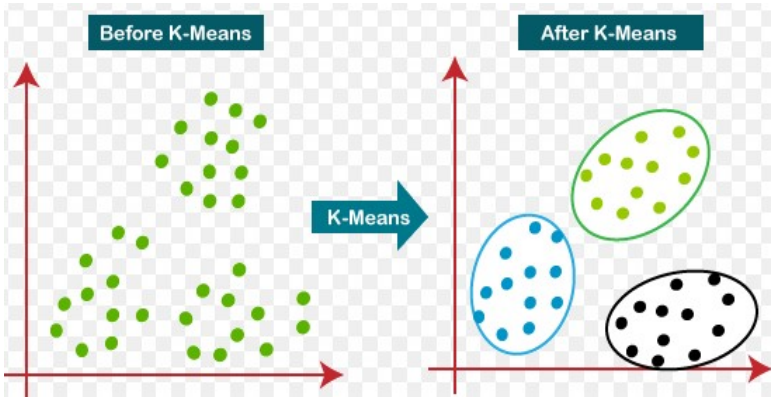

Fig 2 Example of KMC algorithm

Algorithm 1: KMC algorithm

1. Select clusters (k) from gene expression datasets (D).

2. Initialize centres of cluster µ1,… µk

3. Determine cluster centres for k data points and use these points.

4. Take the cluster means and randomly allocate points to clusters.

5. Using the formula below, cluster centres are determined by data points closest to them and computing distance measures to identify missing values.

$$d(i,j) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Where x_i, y_i are Euclidean space's points

6. Put this cluster's data point in there.

7. Recalculate the cluster centres (cluster mean of data points).

8. When there are no more reassignments, stop.

The instances found having missing values were discarded before splitting the dataset. Instances with complete values formed one part while the other had instances that were partially complete or having missing values. The collection of entire occurrences is treated to KMC to build clusters of full instances. As a result, each occurrence is examined individually, and any deficiencies are subsequently filled with potential values. Following formation of clusters, newly inserted instances were checked to ensure that they were clustered in right classes. If the assigned values were found in proper clusters, they were made permanent, and procedure was iterated for further occurrences. Instances in wrong clusters had their next values allocated and compared until their right clusters were found. As a result, pre-processing strategies successfully applied KMC method to successfully improve classification accuracies.

In the preparation stage, the optimal data records (615 cells) were converted into images of 25 x 25 pixels, the cancer gene expressions are stored into memory, and numerical range of data altered from [0, 24248] to [0, 255] based on Equation 6.

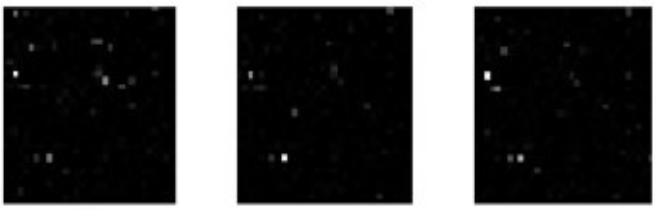

Fig 3 – Outputs of pre-processing phases for a) KIRC b) BRCA and c) LUSC

3.2 Feature selections using AFO algorithm

AFO algorithm was used for successful selection of features. AFO generates fitness functions and feature selections for classification tasks. This work minimizes counts of features, improved classification accuracies, and only most significant RNA-seq properties are selected.

For the FA, the biological and societal traits of real fireflies provided as inspiration [15]. Real fireflies produce a quick, rhythmic flash that serves as a signal of impending danger as well as a help in luring (communicating with) their spouse. The objective function for optimisation of the issue is used by FA to formulate this flashing characteristic. It operates in a similar way to how a firefly's flashing lights do. A firefly group is prompted by the brightness of the light to travel to alluring and intense locations that are designed to produce the best resolution over the desired place.

A number of the firefly traits are normalised by this technique, which may be demonstrated as follows [16]:

(i) Fireflies are drawn to different people regardless of their sex.

(ii) When there are two fireflies nearby, brightness is an important for attractions between them and is exactly proportionate to the appeal. A firefly will randomly reverse direction if it cannot locate a brighter firefly in the area.

The objective function statistically influences firefly's brightness. FA was chosen in this work because of its capability of finding best answers to problems with many objectives. Maximising brightness are proportional to objective functions or brightness or light intensities of firefly's attractions are characterised by encoded goal functions.

a) Source attractiveness and light intensity: According to the inverse square law, the amount of light varies as follows.

$$I(r) = \frac{I_0}{r^2} \quad (2)$$

Where I(r) stands for light intensities at attractiveness r^2

Attractions generated pixels' randomized assignments

b) While intermediates are specified, light intensities can be:

$$I(r) = I_0 \exp(-\gamma r) \quad (3)$$

Where I_0 represents medium's absorption coefficient

c) Singularities are avoided in computations by using Gaussian approximations given below:

$$I(r) = I_0 \exp(-\gamma r^2) \quad (4)$$

The amount of light that the nearby fireflies can see directly affects how appealing a firefly is. By taking into account the potential for variation and randomly altering the allocation, a new solution is achieved. Certain required characteristics are given to pixels in batches.. Hence, a firefly's attractiveness β consists includes:

$$\beta = \beta_0 \exp(-\gamma r^m) \quad (5)$$

Where β_0 is the attractiveness at r=0.

Distances between fireflies i and j ((pixels)) are computed as follows:

$$r_{i,j} = \sqrt{\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2} \quad (6)$$

Where $x_{i,k}$ implies spatial match's kth factor, x_i the kth firefly and d stands for dimensions. AFO used includes adaptation parameter for absorptions and randomizations and create the most effective picture features from the database. These changes increase capabilities of local and global searches by changing parameter values linearly in iterations [16]. Parameters α were computed as:.

$$\alpha(t+1) = \left(1 - \frac{t}{MaxG}\right)\alpha(t) \quad (7)$$

To increase solution precision and convergence time, α adjusts to values with the optimization's distance deviation degrees. Simultaneously, it is recast as follows to enhance population flexibility:

$$\alpha = \alpha_{min} + (\alpha_{max} - \alpha_{min}) \times ||x_i - x_{best}||/L_{max} \quad (8)$$

$$\text{Where } L_{max} = (x_{worst} - x_{best}) \quad (9)$$

Where α_{max} stands for max. characteristics while α_{min}, implies min. Characteristics. In Eq. (9), x_{worst} represents worst generations of individual fireflies t, and L_{max} shows the distances between poorest individuals x_{worst} and global ideal individuals x_{best}. Dispersed fireflies are at longer distances from globally ideal individuals in initial stages where values of $\|x_i - x_{best}\|$ are greater, and $L_{max}$and $(\alpha_{max} - \alpha_{min})$ are constant values. According to Eq. (8), higher values in early stages have greater overall optimizations. Individual fireflies i are drawn towards other fireflies that are brighter due to algorithmic implementation results. Values of $\|x_i - x_{best}\|$ get reduced, which aids in searching best features. Algorithmic convergences are also hastened by considerations of optimum positions whenever values are changed. The step sizes which balance capabilities of algorithmic developments and searches, vary adaptively and dynamically in accordance with distances between fireflies. A unique fitness function was developed in this work based on execution times and accuracies and as:

$$f(x) = \frac{\left({}^{I_d}/_{I_t}\right) \times \left({}^{I_p}/_{P_{init}{}^{i}}\right)}{exp^{-{}^{e_E}/_{e_M}} + H_{accuracy}} \quad (10)$$

where I_d represents dropped image feature counts while m_t represents counts of highly accurate feature sent

I_p stands for image i's pixel.

$P_{init}{}^{i}$ represents initial images.

e_E implies execution times where e_M represents max allowed delays.

$$x_i = x_i + \beta_0 e^{-\gamma r^2}\left(x_j - x_i\right) + \alpha\left(rand - \frac{1}{2}\right) \quad (11)$$

Where $x_i, x_j \ \ are$ distances between two firefly features

Fitness values of characteristic are computed and determined in first generations, whereas pixels counts in batches are generated at random. Selection procedures subsequently choose two fireflies for generating further generations. Fireflies with highest fitness ratings and more brightnesss are selected.

Algorithm 2: AFO

Input: Tumor gene expression dataset

Output: Optimal features

1. Start
2. Objective functions (x), $x = (x1,...,)T$ account for higher accuracies as objective functions
3. Generate initial population of fireflies xi (i = 1, 2, . . . , n), light intensity Ii at xi found using $f(xi)$, set light absorption coefficients γ
4. Estimation of objective functions, AFO computes objective functions to conclude on best values.
5. while (t < MaxGeneration)
6. for i=1:n all n fireflies (features)
7. for j=1:i all n fireflies (features)
8. if ($Ij > Ii$), Move firefly i towards j in d-dimension;
9. end if
10. changed attractiveness based on distances r using exp[$-\gamma r$]
11. computations of fitness using (10) and (11)
12. computing objectives using (6).
13. Estimating new solutions using (3) and correspondingly update light intensities.
14. Use (8) to update the ideal characteristics.
15. end for j
16. end for i

17. Rank features and find current best features
18. end while
19. return optimal features

3.3 Classification using ECNN algorithm

This work's recommended deep learning architectures included sizable amount of learnable parameters in training. On pre-processes original datasets contained 2086 pictures for five different cancer gene expressions. Models get overfitted when large discrepancies between learnable parameters counts and images counts in training sets. When given vast datasets, deep learning models perform better. Jittering, often known as data augmentation, is a popular approach for expanding databases. Data augmentation can expand the dataset size by up to tenfold while training on very little data or 20 times of original sizes or assists in minimizing overfits.

3.3.1 Augmentation process

Increasing the counts of training samples help in avoiding overfitting while maintaining label information. In order to increase the final model's robustness to reflections, magnifications, and mild pixel noises, training sets are additionally subjected to data augmentations which can be done using:

- *Reflection around X axis*
- *Reflection around Y axis*
- *Reflection around X-Y axis*
- *Zooming*

The dataset's picture count increased from 2086 to 10430 after the application of the aforementioned data augmentation techniques, or by a factor of five benefitting training phases of neural networks. Moreover, this also ensures dependability of deep learning architectures is tests along with resistance to data memorizations. Strategies using AI and data mining aim to significantly improve predictions from gene expression databases.

3.3.2 ECNN architecture

In this study, it is suggested that ECNN be used to get findings for pest classification that are more precise for the specified micro array database. ECNN categorizes test data into two classes namely yes or no and obtains higher accuracy.

Input layers

To accurately transfer data to subsequent layers, input layers transform tumour gene images from training images into coherent forms..

Convolutional Layers

The convolution layers analyze the item descriptions in the segmented tumour image data as feature extractors. Convolution layers' neurons are structured into feature maps. A reachable accessible region is associated with a neural neighbourhood in the preceding layer by a number of trainable mean weights in each neuron in a function map [17]. The inputs and

observed weights are combined with a nonlinear activation function to generate the new feature map. Although the neurons in each feature map must have the same weights, various feature maps within the same convolutional layer can have varying weights, allowing for the extraction of a range of characteristics at different places. Specifically, the kth output factor Yk may be calculated as

$$Y_k = f(C_k * x * m_{WE}) \quad (12)$$

The segmented input images are denoted by x, convolution filters are connected to kth feature maps and denoted by Ck. The multiplication operator signifies 2D convolution operators are used in computations of inner products of filter models at locations of fragmented input images, and nonlinear activation functions are denoted by f(•). Nonlinear extractions are possible by the usage of nonlinear activations. Presently, this study uses Rectified Linear Units (ReLUs) [18]. For each fragmented input picture "x," the mean weight value is calculated in equation (13). For each image, a set of random weight values w_i,i=1... n) are offered as input. Following the determination of the mean value for those random weight values, the final mean is multiplied by the input picture that has been broken up.

$$m_{WE} = \frac{W}{\sum_{i=1}^{n} w_i} \quad (13)$$

Where Wrefers to the weight of the single image, $w_i$refersto the random weights.

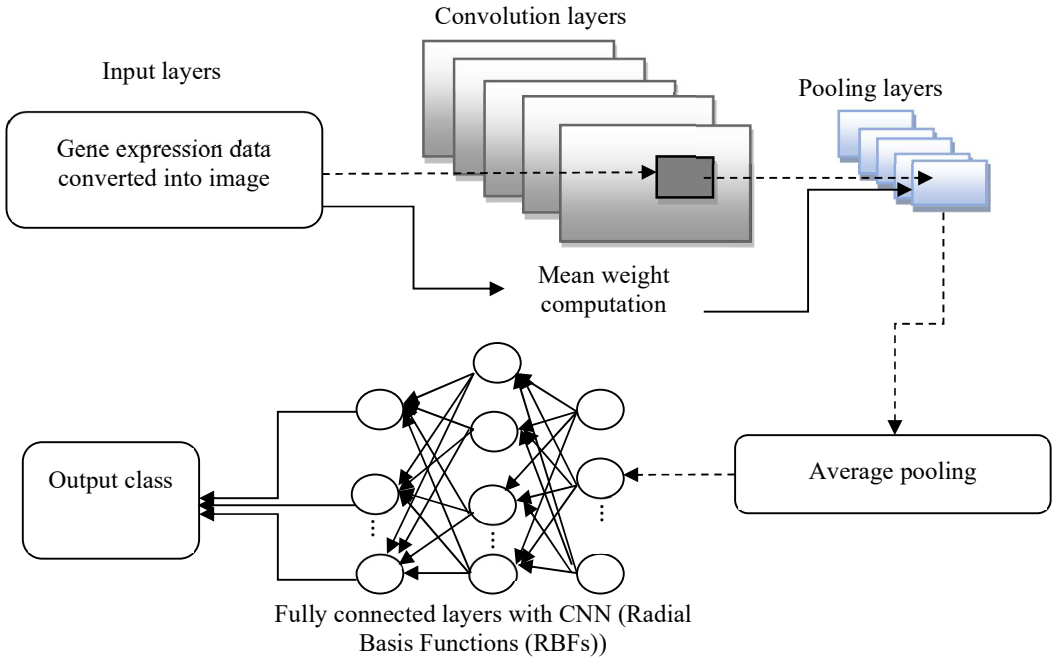

Fig 4 ECNN architecture

Pooling Layers

Pooling layers are used in the feature maps to reduce spatial resolution and establish spatial invariance to input misrepresentation and translation [22]. It is usual practise to multiply

the average of all input values from one layer to the next from a certain picture neighbourhood using average pooling aggregation layers. ECNN architecture is shown in Fig. 4,.

Fully Connected Layers

Convolution and pooling layers are stacked on top of one another for building complicated representations of features. These feature representations are seen by the fully connected layers underneath these layers, which are performing the process as high-level reasoning. For classification issues, the softmax technique combined with a CNN [19] is used. Radial Basic Functions (RBFs) classified convolution layers resulting in increased classification accuracy.

4. Experimental work

The dataset for cancer gene expression from this investigation was released in [40]. It included RNA sequencing results from five different cancer types including UCEC (uterine corpus endometrial carcinoma), BRCA (breast invasive carcinoma), KIRC (kidney renal clear cell carcinoma), and LUAD (lung adenocarcinoma). This dataset's 2,086 rows and 972 columns represent distinct samples, and columns represent RPKM RNA-Seq values for distinct genes. The final column uses counts to encode different forms of cancer: 1D BRCA, 2D KIRC, 3D LUAD, 4D LUSC, and 5D UCEC. Table 1 shows the number of samples collected for each form of cancer.

Table 1 Counts of Samples in Datasets

Tumor type	BRCA	KIRC	LUAD	LUSC	UCEC
Number of samples	878	537	162	240	269

In this part, the effectiveness of the new AFOECNN scheme is evaluated and compared to older methods like the CNN [20] and BPSO+CNN [21] algorithms. The tests are run through Matlab. We evaluate the present and recommended techniques' precision, recall, f-measure, and temporal complexity.

Accuracy

Accuracies of models are computed as total actual classified parameters (T_p+T_n), which are then divided by parameters of total classifications (T_p+T_n+F_p+F_n). The calculation for precision is as follows:

$$\text{Accuracy} = \frac{T_p+T_n}{(T_p+T_n+F_p+F_n)} \quad (14)$$

Where Tp is true positive, Tn is true negative, Fp is false positive and Fn is false negative

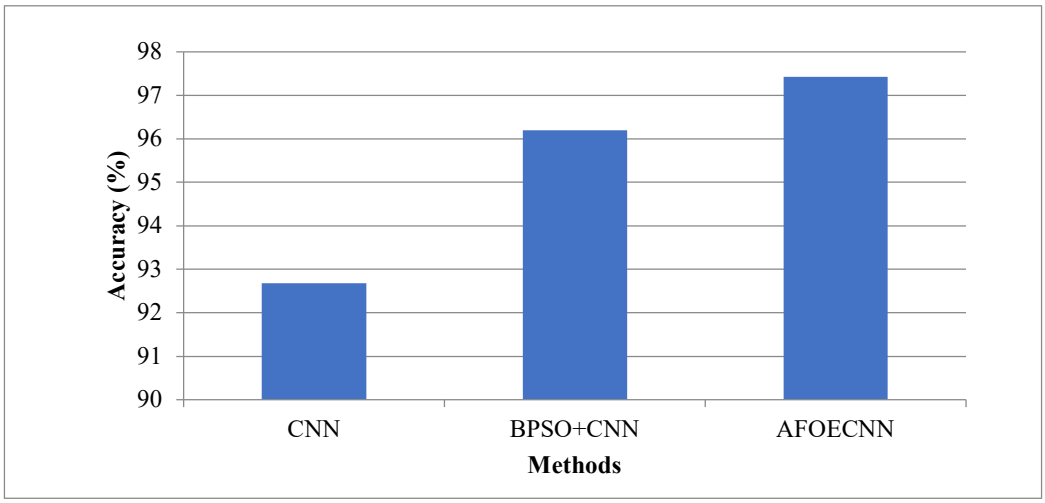

Fig 5 Accuracy

Fig. 5 which depicts evaluations accuracy values for existing and proposed approach where x-axis is formed by approaches while their accuracy values are y-axis values. Conventional methods like CNN and BPSO+CNN algorithms perform less accurately on the presented gene expression dataset, however the new AFOECNN algorithm performs better. Pre-processing is used with the KMC approach to raise classification accuracy. The suggested AFO-based feature selection improves the relevant qualities for better results. According to the findings, the suggested AFOECNN algorithm enhances the efficiency of the gene expression dataset using the optimised technique.

Precision

The precision is calculated as follows:

$$\text{Precision} = \frac{T_p}{T_p + F_p} \tag{15}$$

While precision values can be considered as accuracy or quality assessments, recall values are used to assess completeness or quantity. Accurate algorithms produced more relevant results. The proportion of real positives to all objects labelled as belonging to the positive class in a classification task determines the accuracy of a class.

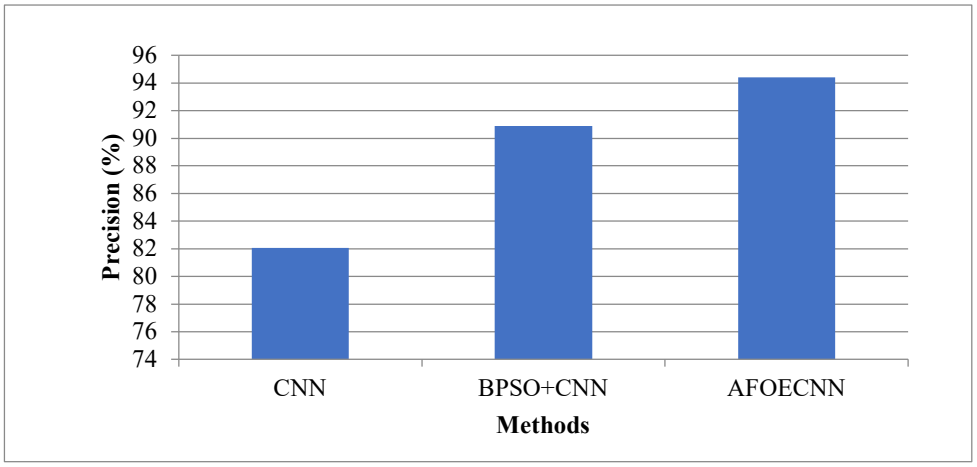

Fig 6 Precision

Fig. 6 which depicts evaluations precision values for existing and proposed approach where x-axis is formed by approaches while their precision values are y-axis values The proposed AFOECNN algorithm outperforms the current CNN and BPSO+CNN techniques in terms of precision. The suggested approach is focused on choosing the more pertinent data. By using the AFOECNN algorithm, diseases may be detected early on and treated. In light of these findings, it can be said that the suggested AFOECNN technique improves the accuracy of the useful characteristics for the illness identification process.

Recall

The calculation of the recall value is done as follows:

$$\text{Recall} = \frac{T_p}{T_p + F_n} \tag{16}$$

The comparison graph is depicted as follows:

Recall values are ratios of relevant document counts found by searches against total document counts found by that search, whereas precision values are ratios of relevant documents counts found by a search to total document counts returned by that search.

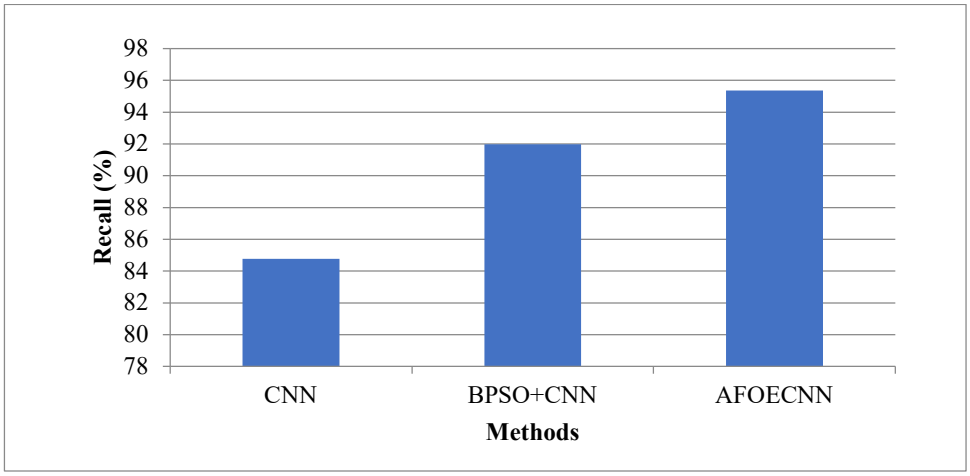

Fig 7 Recall

Fig. 7 which depicts evaluations recall values for existing and proposed approach where x-axis is formed by approaches while their recall values are y-axis values The proposed AFOECNN algorithm outperforms current BPSO+CNN techniques in terms of recall. The proposed method focuses on selecting the most relevant data from the gene expression dataset's many features. Based on these findings, the proposed AFOECNN technique increases the accuracy of the useful characteristics for the sickness identification process.

F-measure

F1-score is defined as:

$$\text{Fscore} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + r} \tag{17}$$

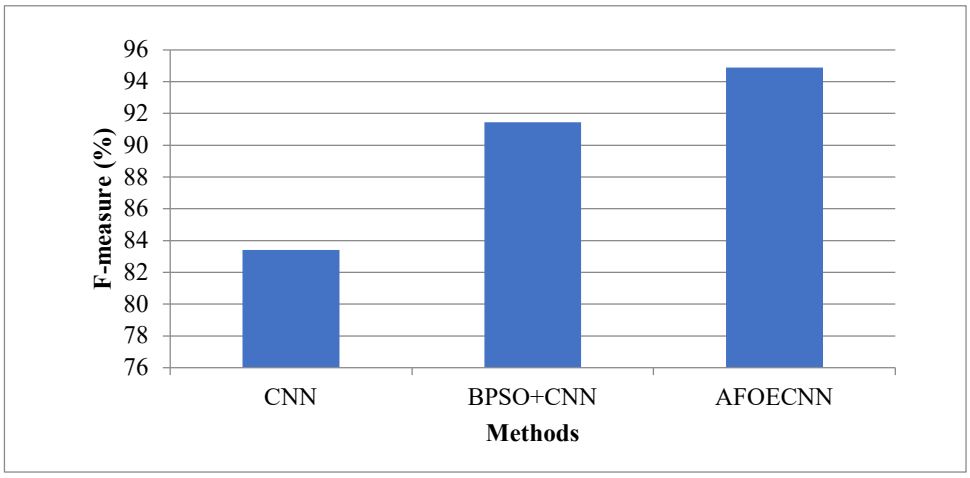

Fig 8 F-measure

F1 scores are balanced assessments of precision and recall values in validations, accounts for both precision and recall equally. Fig. 8 which depicts evaluations F1-scores for existing and proposed approach where x-axis is formed by approaches while F1-scores are y-axis values. The proposed AFOECNN algorithm outperforms the current CNN and BPSO+CNN methods in terms of F-measure when applied to the presented gene expression dataset. The proposed AFOECNN classifier successfully predicted features with an F1 score of 94.88% and no wrongly detected features. To discriminate between impacted and unaffected characteristics, a more relevant feature set is chosen using the AFO model. As a result, the proposed AFOECNN algorithm offers superior illness detection performance and greater disease classification accuracy.

Time complexity

The proposed algorithm is better when it provides lower time complexity

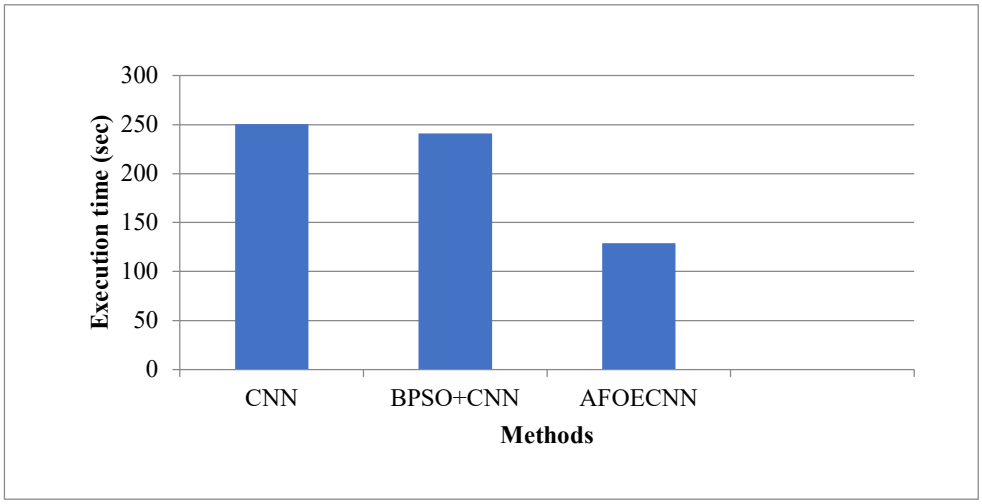

Fig 9 Time complexity

As illustrated in Fig. 9, both the current and proposed techniques are used to evaluate the comparison measure in terms of execution time. The x-axis represents approaches while y-axis are representations of their execution times. Current approaches, such as CNN and

BPSO+CNN, take longer to run on the supplied gene expression dataset than the proposed AFOECNN algorithm.

5. Conclusion

This work suggests AFOECNN approach to enhance pest classification accuracies for supplied datasets. The four main parts of this study are dataset collection, pre-processing, feature selection, and classification. The KMC approach is initially used to collect and pre-process the cancer genes. The features are chosen using the AFO algorithm, which selects the top features quickly. The AFO algorithm then provides relevant and useful features that are used in real-time applications. It used RNA-Seq to choose the best features before converting them to two-dimensional pictures. data augmentations increased original dataset sizes. The classifications were executed using ECNN technique, which improved detection performances. The AFOECNN model suggested here enhances classification accuracy. In terms of accuracy, precision, recall, f-measure, and temporal complexity, the experimental findings show that the proposed AFOECNN approach beats state-of-the-art techniques. The Generative Adversarial Neural (GAN) network may be employed in future studies, and the deep learning architecture may be enhanced utilising tools like as AlexNet, Vgg-16, and Google-Net.

References

1. Tabakhi, Sina, et al. "Gene selection for microarray data classification using a novel ant colony optimization." *Neurocomputing* 168 (2015): 1024-1036.
2. Vadapalli, Sreya, et al. "Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine." *Briefings in bioinformatics* 23.5 (2022).
3. Umair Shafique, Fiaz Majeed, Haseeb Qaiser, and Irfan Ul muftaja "Data mining in healthcare for heart disease" International Journal of Innovation and Applied Studies",ISSN 2028-9324 Vol.4,pp 1313-1322, Mar 2015
4. Bhardwaj, Brijesh Kumar, and Saurabh Pal. "Data Mining: A prediction for performance improvement using classification." *arXiv preprint arXiv:1201.3418* (2012)
5. Sheikhpour, Razieh, Mehdi Agha Sarram, and Robab Sheikhpour. "Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer." *Applied Soft Computing* 40 (2016): 113-131.
6. Aalaei, Shokoufeh, et al. "Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets." *Iranian journal of basic medical sciences*19.5 (2016): 476
7. Singh, Shailendra. "A novel algorithm to preprocess cancerous gene expression dataset for efficient gene selection." *2017 2nd International Conference for Convergence in Technology (I2CT)*. IEEE, 2017.
8. Chowdhury, Shanta, Xishuang Dong, and Xiangfang Li. "Recurrent neural network based feature selection for high dimensional and low sample size micro-array data." *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019.

9. Danaee, Padideh, Reza Ghaeini, and David A. Hendrix. "A deep learning approach for cancer detection and relevant gene identification." *Pacific symposium on biocomputing 2017*. 2017.
10. Sawhney, Ramit, Puneet Mathur, and Ravi Shankar. "A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis." *Computational Science and Its Applications–ICCSA 2018: 18th International Conference, Melbourne, VIC, Australia, July 2-5, 2018, Proceedings, Part I 18*. Springer International Publishing, 2018.
11. Zhang, Long, Linlin Shan, and Jianhua Wang. "Optimal feature selection using distance-based discrete firefly algorithm with mutual information criterion." *Neural Computing and Applications* 28 (2017): 2795-2808.
12. Liu, Kui, et al. "Breast cancer classification based on fully-connected layer first convolutional neural networks." *IEEE Access* 6 (2018): 23722-23732.
13. Pfeffer, Maximilian Achim, and Sai Ho Ling. "Evolving Optimised Convolutional Neural Networks for Lung Cancer Classification." *Signals* 3.2 (2022): 284-295.
14. Mohamad, Ismail Bin, and Dauda Usman. "Standardization and its effects on K-means clustering algorithm." *Research Journal of Applied Sciences, Engineering and Technology*6.17 (2013): 3299-3303
15. Liu, Jingsen, et al. "A dynamic adaptive firefly algorithm with globally orientation." *Mathematics and Computers in Simulation* 174 (2020): 76-101.
16. Liu, Changnian, et al. "Adaptive firefly optimization algorithm based on stochastic inertia weight." *2013 Sixth International Symposium on Computational Intelligence and Design*. Vol. 1. IEEE, 2013
17. Lopez-Garcia, Guillermo, et al. "Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data." *PloS one* 15.3 (2020): e0230536.
18. Yu, D., Wang, H., Chen, P., & Wei, Z. (2014). Mixed pooling for convolutional neural networks. *In Proceedings of the 9th International Conference on Rough Sets and Knowledge Technology,* pp. 364–375
19. Tools and Machine Learning Algorithms for Predicting Depression, Anxiety, and Stress: A Literature Review, "International Journal of Research Publication and Reviews", Volume- 4, Issue -3, pp 1108-1114, March 2023, 5.536.
20. Kong, Yunchuan, and Tianwei Yu. "A deep neural network model using random forest to extract feature representation for gene expression data classification." *Scientific reports* 8.1 (2018): 16477.
21. Elbashir, Murtada K., et al. "Lightweight convolutional neural network for breast cancer classification using RNA-seq gene expression data." *IEEE Access* 7 (2019): 185338-185348.
22. Khalifa, Nour Eldeen M., et al. "Artificial intelligence technique for gene expression by tumor RNA-Seq data: a novel optimized deep learning approach." *IEEE Access* 8 (2020): 22874-22883