ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

A Novel feature extraction based neural network model on high dimensional somatic cancer prediction

P R Sudha Rani[1], Dr. K R Ramya[2]
[1]Research Scholar, [2]Professor
[1,2] Department of CSE, Acharya Nagarjuna University (University College of Engineering and Technology) Nagarjuna Nagar-522510, Guntur, A.P, India
[1]Email: sudharani.p@gmail.com

ABSTRACT

Due to its high performance and processing speed, the ensemble classifier is an excellent classification model for real-time applications. Standard extreme learning methods cannot anticipate the error rate since most conventional neural network models use a static weight selection. Medical problems can be predicted using a new weighted extreme learning machine (WELM). It is the fundamental purpose of the weighted extreme learner to define high-dimensional data for disease prediction. To improve cancer prediction, the proposed ensemble approach often uses high-dimensional data. A number of ensemble learning models, including random forest, neural networks, ACO+NN and PSO+NN, were used to evaluate the WELM model proposed in this paper's publication. Medical datasets, including liver, diabetic, ovarian, and DLBCL-Stanford, are used to assess test outcomes. Medical databases benefit from the WELM's computing efficiency, as measured by its true positive rate, its error rate, and its accuracy.

I. INTRODUCTION

Using supervised learning, a sample of data is first sorted into a pre-existing set of categories. New samples are assigned to an existing class or label by means of a classifier[8-11]. Numerous patterns, such as high-dimensional data analysis and natural language processing as well as handwriting recognition and computer biology can all benefit from the classification of data. Other applications for data classification include drug design and the identification of disease in patients. During the training phase, the classification algorithm generates a classifier that examines the training samples and their associated attributes and labels. If you have access to class labels in advance, this is termed supervised learning. The classifier's performance is evaluated again using the test samples in the second stage. The training samples are used to generate the test samples, which are then drawn at random from the entire dataset. The test result is evaluated using a variety of performance metrics. Filtering strategies are used to pick a subset of features prior to selecting a classifier. embedded approaches select functions that are typically unique to each learner within the integrated process.

According to semi-supervised feature selection strategies, function importance is evaluated based on the ability to retain specified information qualities, such as variance or locality preservation capacity. Because of the usage of marked data, semi-supervised feature selection strategies typically produce less

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

efficient outcomes than unsupervised feature selection methods[3]. To be sure, supervised procedures for picking features require appropriate marked information that requires extensive knowledge and is expensive to acquire Label data and data allocation information or the local mix of both marked and unlabeled data are used to determine the function's relevance in semi-supervised decision procedures [4]. Semi-supervised strategies for picking features have not yet been extensively studied. The majority of traditional models go into great detail about semi-supervised methods of feature selection, classifying them according to two unique criteria, summarizing the data they give, and outlining the benefits and drawbacks of each. There has never before been a comprehensive study on semi-supervised function choosing strategies that categorizes them from two unique perspectives. As a result, semi-supervised education relies heavily on unlabeled data rather than marked information. Semi-supervised learning[6] necessitates the use of smoothness criteria such cluster assumption[5] and multiple hypothesis. Microarray data analytics has been studied for the past ten years using machine learning approaches. Every one of these studies is aiming to produce physiologically important interpretations of large data sets, which will be useful for further research. Studies aiming to find genes with important biological ties to classifications necessitate feature selection approaches in order to ensure researchers comprehend their data in high-dimensional scenarios like this one. Filtering is the most common method for dividing input characteristics. It is possible to considerably improve the classification process by using a proper feature selection method. Different statistical assessments are used to identify the subdivision of features along with a suitable predicting ability in this procedure. First, a statistical evaluation is conducted, and the final score for each feature is calculated. Pre-processing for classification is incomplete without the feature selection procedure. In order to avoid the problem of dimensionality curse, the selection of features is regarded the most critical process[7-10]. In single-layer feed-forward networks, the extreme learning machine is now considered the learning algorithm. It is possible to use a variety of extreme learning devices in a variety of biomedical applications. As a result of microarray data's infeasibility and complexity, predicting illness status is difficult. In the process of converting the gene's DNA sequence into the necessary mRNA sequences, gene expression is considered an important mechanism. Countless genes have their levels of expression regulated by this complex molecule. Cluster analysis is the most effective method for examining the complexities of medical conditions. In addition to this, it will improve the method of prognosis. It is necessary to select a method for selecting particular parameters in all traditional clustering methods[11-14]. To put it another way, classification is a method for sorting items of interest into discrete groups according to previously established categories. Non-linear feature selection based semi-supervised learning model on high dimensional datasets to improve the error rate and true positivity are shown in this part of the paper.

II. RELATED WORKS

An tremendous amount of data may be collected quickly and easily using a DNA microarray analysis, which generates thousands of measurements of gene expression. More objective, accurate, and dependable than standard diagnostic methods are gene expression profiles.[15] However, when using this data for data mining,

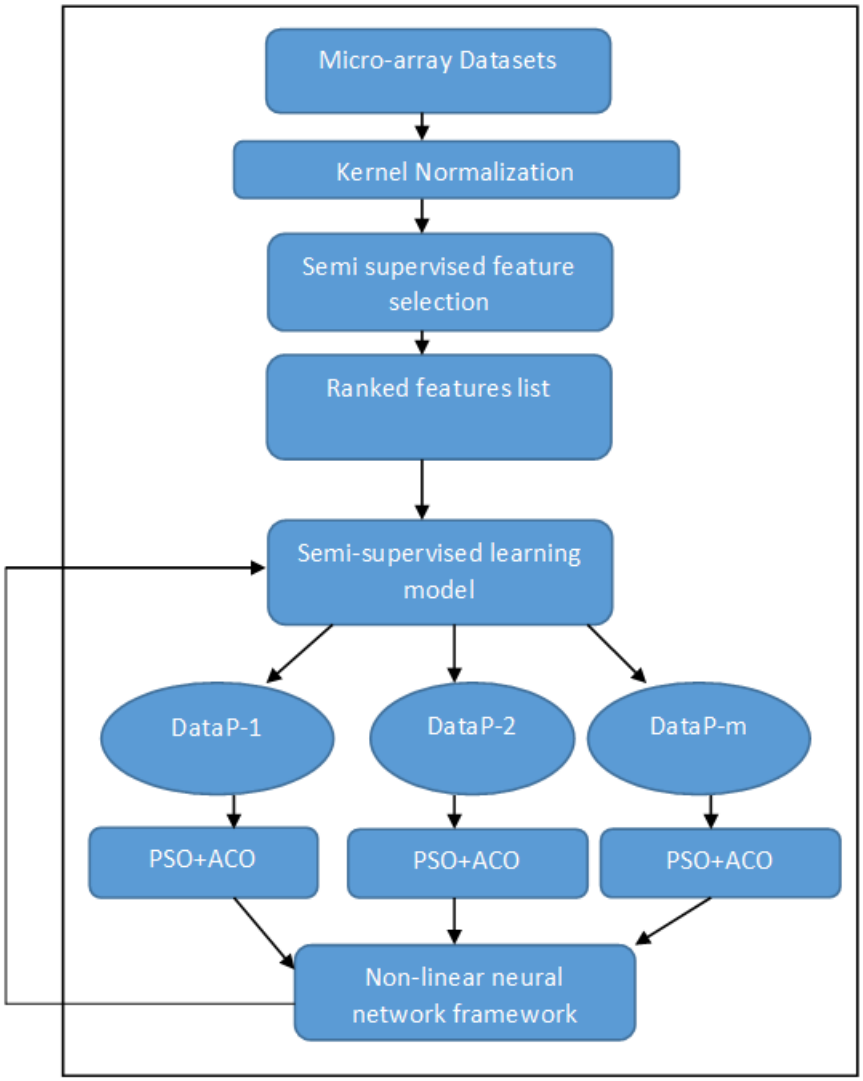
ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

there are a number of issues. The sheer scale of these databases is one of their most significant flaws. As a result, there is not enough data to train and test the models. Despite the vast number of genes, only a handful are significant and the rest is either redundant, noisy, or less important. The presence of these genes lowers classification and diminishes prediction performance, while increasing the rate of errors. We require human experts to pick a collection of data most importantly. When it comes to unlabeled information, though, it's generally copious, straightforward, and inexpensive to obtain. A few issues, however: for unlabeled data or a test set, the procedure of creating data may differ from that for labelled data or an instructor's training data. Using "semi-supervised teaching" methodologies, researchers have been able to build more robust classifications that can be utilised beyond the "out of pockets" level of research (i.e. test sets). Our model uses both tagged and unlabeled data to learn. A learning parameter[16] controls the rate at which the brain learns from unlabeled material. To estimate the learning parameter, we also present a Bayesian technique to maximise marginal probability. Classifications of this type were chosen since they fall within a variety of categorization categories and are well-known for their accuracy[17-20]. For this specific form of data gathering in training, multiple categorization algorithms are used, and one is obviously more effective than the others. The J48 (decision tree classification) classifier is one of the most often used. Classifiers IB1 and IBk, known as "lazy" classifiers, construct their classification models by looking at the data from the closest neighbours. Support vector machines, a concept found in support vector machines, are used in SMO's classification technique. The Bayes classifier group includes NaiveBayes, which employs calculator classes for categorization. NaiveBayes[21-22] and if-then rules[21-22] are used in NNge, a rules-based system that shares certain characteristics with the Bayes classification.

3. Proposed Model

The overall architecture of the proposed model is represented in fig 1. Initially, each microarray gene disease dataset is processed to find the synonym of the gene feature for efficient gene-symbol to gene-name mapping.

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019



Microarray training data is pre-processed using data transformation functions to remove any volatility in data distribution. Data transformation functions using kernels are employed to standardise training data for clustering and wrapper feature ranking in the mapper phase of the proposed study.

Kernel based Data Pre-processing

Input : Training microarray dataset D, F(D): Feature space of D, Max similarity MaxSim[], Threshold T.

Output: Kernel Filtering or Transformed data KD.

1. Read input data D.
2. For each pair of feature F[i],F[j] in feature space F(D)
3. Do
4. Apply Kernel transformation on I as
5. $$\text{GeneKernelTransform}(F[i],F[j]) = \frac{1}{1+\eta^2 / \cos(\max(\sigma^2_{F[i]}, \sigma^2_{F[j]})}$$

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

Procedure:

Where $\eta = 2 * \sum \text{Feat}[i].\text{Feat}[j] - \sum (\text{Feat}[i] + \text{Feat}[j])^2$

- If($\text{GeneKernelTransform}(F[i], F[j]) > T$)
- Then
- Normalize Feat[i] and Feat[j] within $[0, \text{GeneKernelTransform}(\text{Feat}[i], \text{Feat}[j])]$ using Min-max normalization as KD
- Else
- Normalize F[i] and F[j] within [0,1] using Min-max normalization as KD

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

- End if
-
- Compute Similarity between gene symbol and gene name using the similarity measure.
- $$\text{Gsim}[] = \max\{\frac{1}{3}\{\alpha(\frac{1}{|\text{FS}(D,i)|} + \frac{1}{|\text{GD}[j]|}) + (1 - \frac{t}{\beta})\}, \frac{|\text{FS}(D,i) \cap \text{GD}[j]|}{|\text{FS}(D,i) \cap \text{GD}[j]| + \alpha|\text{FS}(D,i) - \text{GD}[j]| + \beta|\text{GD}[j] - \text{FS}(D,i)|}\}$$
- Done

In this case, the values within the provided range are normalised using minimum and maximum scaling. High dimensional datasets are cleaned using this method.

Wrapper feature ranking based non-linear neural network classification

T-statistics, SAM, and SNR are used to rate the clustered features of k-means on microarray datasets using traditional feature selection techniques. Using the wrapper technique to choose appropriate genes from the high-dimensional feature space is a major drawback of these feature ranking measures. These ranking metrics integrate classification accuracy and true positive rate on the specified features (> 50) using t-test, SAM and SNR measurements. Each attribute should be assigned a weighted value based on the maximum weights of (1), (2), and (3). The T-statistical weighting metric is used to detect the variance in gene features based on the standard deviation of the class labels. The ratio of the class label to the maximum standard deviation is called the class-to-standard deviation ratio.

$$W1 = \frac{\mu_P - \mu_N}{\sqrt{\max\{\sigma_P^2/|P|, \sigma_N^2/|N|\}}} \quad \text{-----(1)}$$

Difference in class label to sum of positive and negative gene disease class standard deviation is referred to as the class difference ratio. The genes with the highest signal-to-noise ratio are given the most weight in this data classification.

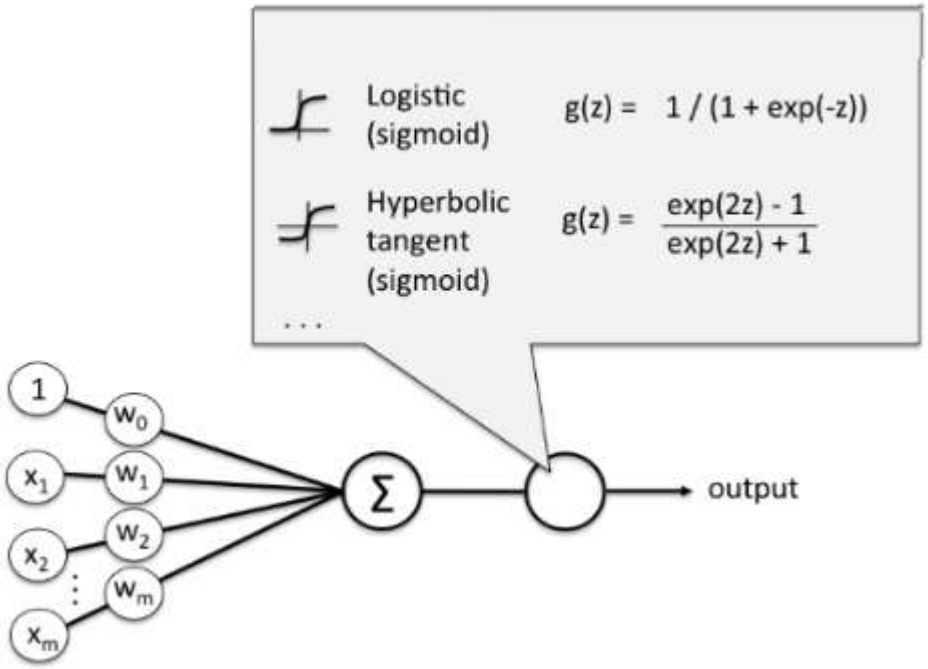
ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

$$W2 = HSNR = \frac{|\mu_i - \mu_j|}{2(\sigma_P + \sigma_N)} \quad \text{-----(2)}$$

where μ_P and σ_P are the mean and standard deviation of the cluster positive class samples μ_N and σ_N are the mean and standard deviation of the cluster negative class samples.

$$W3 = MCTSNR = Max\{Correlation(ClusterFeatures : CF), \frac{\mu_P - \mu_N}{\sqrt{\max\{\sigma_P^2/|P|, \sigma_N^2/|N|\}}}, \frac{|\mu_i - \mu_j|}{2(\sigma_P + \sigma_N)}\} \quad \text{-----(3)}$$

It is The goal is to maximise feature correlation, hybrid t-test, and hybrid SNR ratios. This ranking metric is used to identify the binary class in each cluster that performs at its peak.



4. Experimental Results

Microarrays from the biomedical library were used to test the model's performance in existing models. Table 1 provides a summary of the data sets used in the experiment. The experimental results use 10% of the training data as test data for performance evaluation. The genuine positive rate and accuracy of huge datasets are improved using the proposed selection-based ensemble approaches. Therefore, each cross validation tends to have a higher degree of accuracy than a standard ensemble classification model, because the suggested model uses all of the training data. The results of the experiments show that the suggested classification of the ensemble enhances the true positive and negative rate in general.The suggested model's key advantage is that it reduces the error rate on high-dimensional features.

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

Micro array Datasets	Gene sets	Data-Type
Prostate	2136	Continuous/Numeric
Lymphoma	5000	Continuous/Numeric
DLBCL-Stanford	4000	Continuous/Numeric
Breast cancer	24481	Continuous/Numeric
Leukemia	7129	Continuous/Numeric

Table. 1 Datasets and Its Characteristics

Using the proposed methodology, the true positive rate and accuracy of high-dimensional microarray datasets can be improved significantly. As a result, each cross validation prediction is more accurate than in standard ensemble classification models because the proposed model employs the complete training data set for decision patterns.

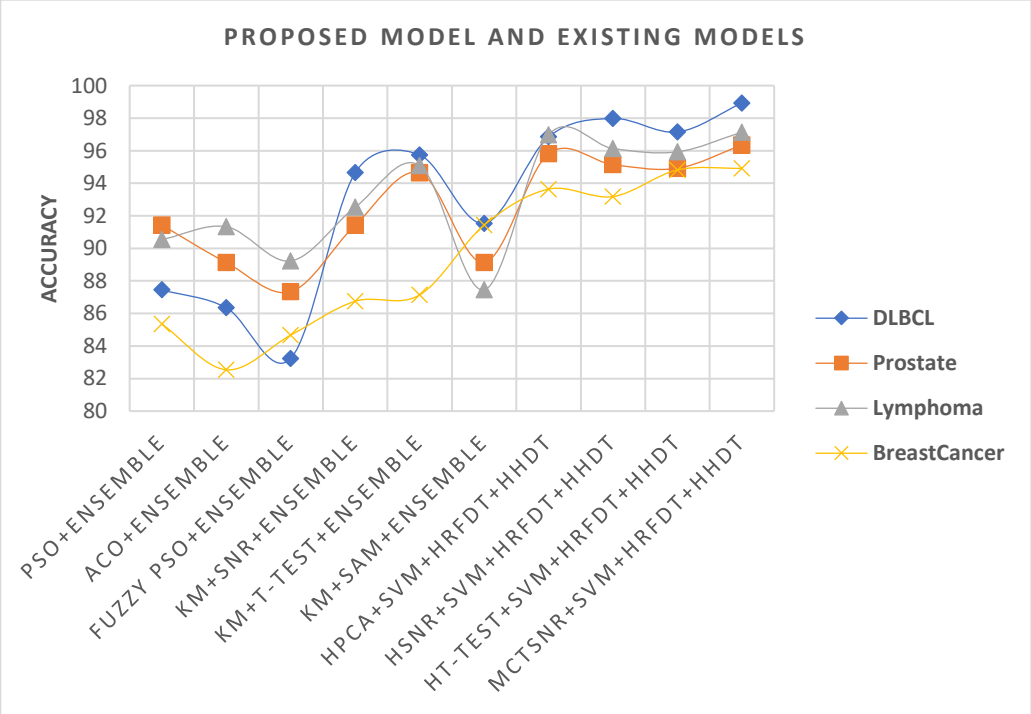


Figure 2:Comparison of proposed model to existing models on average feature seleciton

CONCLUSION

Deep neural networks with weighted functions are used to identify the most important feature sets from a wide feature space.. Using logistic and weighted functions to optimise the deep neural network's weights allows it to effectively classify huge datasets with high dimensionality. An algorithm capable of identifying trustworthy illness candidates utilising current gene-disease connections that can be confirmed through biological

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

experimentation is therefore crucial for effectively resolving these issues As a solution to these challenges, a semi-supervised learning approach based on non-linear feature selection is proposed. The feature space is divided into k-correlated features using a hybrid correlation-based wrapper technique. For the first time, a deep neural network architecture for disease prediction has been devised and implemented. In terms of genuine positivity and receiver operating characteristics, the current model outperforms the previous models (ROCs).

1. References

2. [1]M. Abd-Elnaby, M. Alfonse, and M. Roushdy, “Classification of breast cancer using microarray gene expression data: A survey,” Journal of Biomedical Informatics, vol. 117, p. 103764, May 2021, doi: 10.1016/j.jbi.2021.103764.

3. [2]M. Abdulla and M. T. Khasawneh, “G-Forest: An ensemble method for cost-sensitive feature selection in gene expression microarrays,” Artificial Intelligence in Medicine, vol. 108, p. 101941, Aug. 2020, doi: 10.1016/j.artmed.2020.101941.

4. [3]E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, “Feature selection methods on gene expression microarray data for cancer classification: A systematic review,” Computers in Biology and Medicine, vol. 140, p. 105051, Jan. 2022, doi: 10.1016/j.compbiomed.2021.105051.

5. [4]O. A. Alomari et al., “Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators,” Knowledge-Based Systems, vol. 223, p. 107034, Jul. 2021, doi: 10.1016/j.knosys.2021.107034.

6. [5]N. K. Berry, R. J. Scott, P. Rowlings, and A. K. Enjeti, “Clinical use of SNP-microarrays for the detection of genome-wide changes in haematological malignancies,” Critical Reviews in Oncology/Hematology, vol. 142, pp. 58–67, Oct. 2019, doi: 10.1016/j.critrevonc.2019.07.016.

7. [6]A. Dabba, A. Tari, S. Meftali, and R. Mokhtari, “Gene selection and classification of microarray data method based on mutual information and moth flame algorithm.,” Expert Systems with Applications, vol. 166, p. 114012, Mar. 2021, doi: 10.1016/j.eswa.2020.114012.

8. [7]M. Ghosh, S. Begum, R. Sarkar, D. Chakraborty, and U. Maulik, “Recursive Memetic Algorithm for gene selection in microarray data,” Expert Systems with Applications, vol. 116, pp. 172–185, Feb. 2019, doi: 10.1016/j.eswa.2018.06.057.

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

9. [8]M. A. Hambali, T. O. Oladele, and K. S. Adewole, “Microarray cancer feature selection: Review, challenges and research directions,” International Journal of Cognitive Computing in Engineering, vol. 1, pp. 78–97, Jun. 2020, doi: 10.1016/j.ijcce.2020.11.001.
10. [9]S. Karimi and M. Farrokhnia, “Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: Combining data dimension reduction and variable selection technique,” Chemometrics and Intelligent Laboratory Systems, vol. 139, pp. 6–14, Dec. 2014, doi: 10.1016/j.chemolab.2014.09.003.
11. [10]D. Kundnani and F. Storici, “FeatureCorr: An R package to study feature correlations aided with data transformation for sequencing and microarray data,” Software Impacts, vol. 10, p. 100144, Nov. 2021, doi: 10.1016/j.simpa.2021.100144.
12. [11]R. Kundu, S. Chattopadhyay, E. Cuevas, and R. Sarkar, “AltWOA: Altruistic Whale Optimization Algorithm for feature selection on microarray datasets,” Computers in Biology and Medicine, vol. 144, p. 105349, May 2022, doi: 10.1016/j.compbiomed.2022.105349.
13. [12]L. Meenachi and S. Ramakrishnan, “Metaheuristic Search Based Feature Selection Methods for Classification of Cancer,” Pattern Recognition, vol. 119, p. 108079, Nov. 2021, doi: 10.1016/j.patcog.2021.108079.
14. [13]P. Mishra and N. Bhoi, “Cancer gene recognition from microarray data with manta ray based enhanced ANFIS technique,” Biocybernetics and Biomedical Engineering, vol. 41, no. 3, pp. 916–932, Jul. 2021, doi: 10.1016/j.bbe.2021.06.004.
15. [14]F. Morais-Rodrigues et al., “Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression,” Gene, vol. 726, p. 144168, Feb. 2020, doi: 10.1016/j.gene.2019.144168.
16. [15]E. Nazari, M. Aghemiri, A. Avan, A. Mehrabian, and H. Tabesh, “Machine learning approaches for classification of colorectal cancer with and without feature selection method on microarray data,” Gene Reports, vol. 25, p. 101419, Dec. 2021, doi: 10.1016/j.genrep.2021.101419.
17. [16]H. F. Ong, N. Mustapha, H. Hamdan, R. Rosli, and A. Mustapha, “Informative top-k class associative rule for cancer biomarker discovery on microarray data,” Expert Systems with Applications, vol. 146, p. 113169, May 2020, doi: 10.1016/j.eswa.2019.113169.
18. [17]S. A. B. P, C. S. R. Annavarapu, and S. Dara, “Clustering-based hybrid feature selection approach for high dimensional microarray data,” Chemometrics and Intelligent Laboratory Systems, vol. 213, p. 104305, Jun. 2021, doi: 10.1016/j.chemolab.2021.104305.
19. [18]S. P. Potharaju and M. Sreedevi, “Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance,” Clinical Epidemiology and Global Health, vol. 7, no. 2, pp. 171–176, Jun. 2019, doi: 10.1016/j.cegh.2018.04.001.
20. [19]M. Rostami, S. Forouzandeh, K. Berahmand, M. Soltani, M. Shahsavari, and M. Oussalah, “Gene selection for microarray data classification via multi-objective graph theoretic-based method,” Artificial Intelligence in Medicine, vol. 123, p. 102228, Jan. 2022, doi: 10.1016/j.artmed.2021.102228.
21. [20]S. Sarbazi-Azad, M. Saniee Abadeh, and M. E. Mowlaei, “Using data complexity measures and an evolutionary cultural algorithm for gene selection in microarray data,” Soft Computing Letters, vol. 3, p. 100007, Dec. 2021, doi: 10.1016/j.socl.2020.100007.

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

22. [21]S. Sayed, M. Nassef, A. Badr, and I. Farag, “A Nested Genetic Algorithm for feature selection in high-dimensional cancer Microarray datasets,” Expert Systems with Applications, vol. 121, pp. 233–243, May 2019, doi: 10.1016/j.eswa.2018.12.022.

23. [22]M. Toğaçar, B. Ergen, and Z. Cömert, “Detection of lung cancer on chest CT images using minimum redundancy maximum relevance feature selection method with convolutional neural networks,” Biocybernetics and Biomedical Engineering, vol. 40, no. 1, pp. 23–39, Jan. 2020, doi: 10.1016/j.bbe.2019.11.004.