Disease analysis using Convolutional Neural Network based on gene patterns

N.M.K. RamalingamSakthivelan[1], A Amirthavarshini[2], SP Jayabharathi[3], M Kavimani[4]

[1]Associate Professor, [2,3,4]Students

Department of Computer Science and Engineering, Paavai Engineering College, Namakkal.

Abstract— The DNA microarray technology has revolutionized biological research by enabling simultaneous measurement of the expression levels of thousands of genes in a single experiment. Gene expression profiles, representing the molecular state of a cell, have the potential to serve as a medical diagnosis tool. Disease classification using gene expression data can address fundamental issues related to diagnosis and discovery. Various methods have been proposed in recent years with promising results, but several issues still need to be addressed. This project presents a comprehensive approach to disease classification using pattern similarity search, particle swarm optimization, and convolutional neural network classification. Evaluation time, classification accuracy, and the ability to reveal biologically meaningful gene information are used to estimate the effectiveness of the proposed methods. Our multiclass classification method is applied to diagnose several diseases, including cancer (lung, blood, breast, and skin), determine their severity levels, and prescribe appropriate medicine. Experimental results show improved classifier performance through graphs with increased accuracy.

Key words— Deep learning, CNN classifier, Cancer, Gene expression, Accuracy, Severity levels, clustering, PSO algorithm.

I. INTRODUCTION

Gene-based clustering poses unique challenges due to the special characteristics of gene expression data and the specific requirements of the biological domain. Cluster analysis is often the first step in data mining and knowledge discovery, and the purpose of clustering gene expression data is to reveal natural data structures and gain initial insights into data distribution. Therefore, a good clustering algorithm should rely as little as possible on prior knowledge, which is usually not available before cluster analysis. Another challenge is that gene expression data often contain a significant amount of noise due to the complex procedures of microarray experiments. As a result, clustering algorithms for gene expression data should be capable of extracting useful information from a high level of background noise. In addition, gene expression data are often "highly connected," and clusters may intersect with each other or be embedded one in another. Hence, algorithms for gene-based clustering should be able to effectively handle this situation. Finally, users of microarray data may be interested not only in the clusters of genes but also in the relationships between the clusters and the relationships between the genes within the same cluster. A clustering algorithm that can partition the data set and provide graphical representation of the cluster structure would be preferred by biologists. Overall, gene-based clustering is an open problem, and new algorithms need to be developed to overcome these challenges and provide more insights into gene expression data.

II. LITERATURE SURVEY

Laura Judith Marcos studies the necessary data to fully investigate host-microbiome relationships and how they relate to the onset and course of numerous complicated illnesses. To fully utilize the information from these biological datasets, better data-analytical tools are required, taking into account the unique characteristics of microbiome data. Here, we examine the cutting-edge ML techniques.

Yinan Zhao focused on identification of useful genes from multiple microarrays for diagnosis based on machine learning methods.

Raja Krishnamoorthi study Human malignancies like oral squamous cell carcinoma (OSCC) are rather prevalent. However, nothing is known about its pathophysiology or prognosis. In the current study, we sought to investigate the most important DEGs and how well they predicted OSCC outcome. The Gene Expression Omnibus (GEO) database's several microarray datasets were combined to find DEGs between OSCC.

II. EXISTING SYSTEM

Existing system uses model-based clustering, in which data points in different clusters were generated by using different probability distributions. The number of clusters of given datasets is identified by estimating the parameters using maximum likelihood (ML) or expectation-maximization (EM). The major flaw with algorithms is that they have a slow convergence rate and are sensitive to the initial parameters. K-medoids and K-means (KM) are one of the efficient partition-based that solve well-known clustering problem.

III. PROPOSED SYSTEM

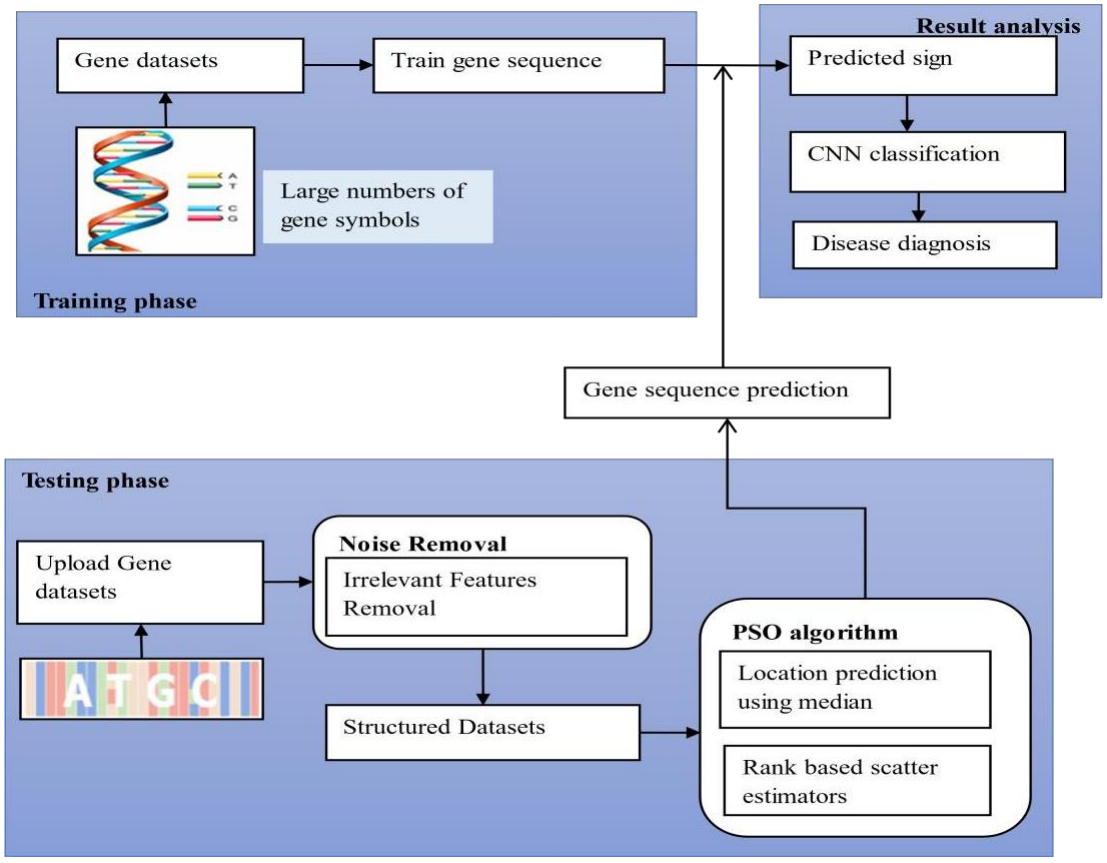
Microarray technology permits simultaneous study of genes comprising a large part of genome. It uses the supervised learning to classify and predict diseases, based on the gene expressions.

Known sets of data will be used to train the deep learning protocols to categorize diseases according to their gene patterns. CNN method provides the information regarding the efficiency of the machine learning techniques.

In response to the rapid development of DNA Micro array technology, classification methods and gene selection techniques are being computed for better use of classification algorithm in microarray gene expression data. The efficiency of classification depends on the type of kernel function used for classification.

IV. ARCHITECTURE

Here first the datasets are uploaded from microarray database and train the gene with algorithms and features are selected.



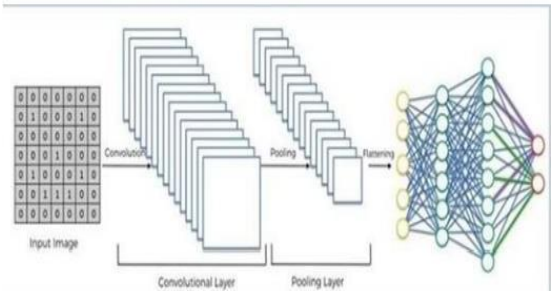
V. ALGORITHMS

PSO ALGORITHM

The Particle Swarm Optimization (PSO) algorithm is an iterative approach for determining the maximum likelihood or maximum a posteriori estimates of parameters in statistical models that involve unobserved latent variables. The PSO iteration involves two steps: the expectation of the log-likelihood is computed using the current estimate for the parameters, and the parameters that maximize the expected log-likelihood are determined based on the P best. It is used to determine the distribution of latent variables in the next G best step. The PSO algorithm is typically used in cases where direct solutions to equations involving latent variables are not possible. The PSO algorithm allows for the analysis of data coverage before clustering and proposes an algorithm that modifies the nearest centroid sorting and transfer algorithms for spatial median clustering. The algorithm involves two phases: transferring an object from one cluster to another, and amalgamating the single member cluster with its nearest cluster. All possible transfers and amalgamations are tested to determine which will improve the value of the clustering criterion.

CNN CLASSIFIER

Convolutional Neural Network represent feed-forward neural networks which encompass diverse combos of the convolutional layers, max pooling layers, and completely related layers and Take advantage of spatially correlation by way of way of imposing a pattern among neurons of adjacent layers. CNN has five layers which is used for processing the image. (1). feature selection layer (2). Classification layer (3). Convolutional layer (4) Depth layer (5) Pooling layer.



VI. MODULES

1.DATASET ACQUISITION

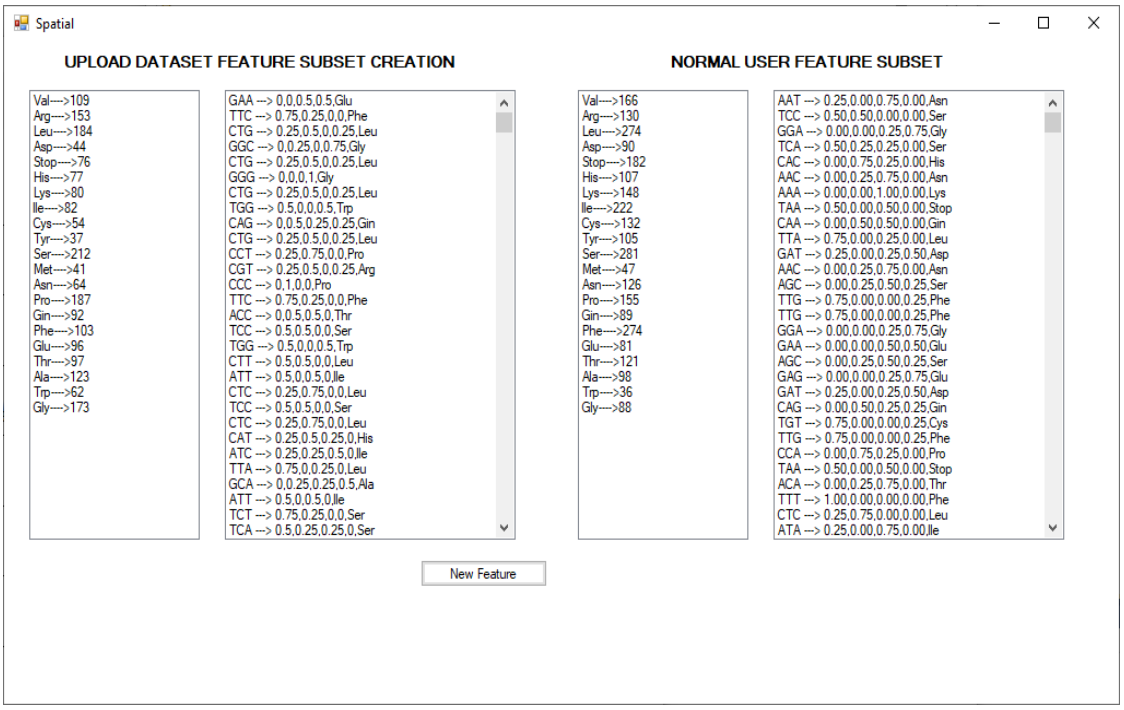
A microarray database is a repository containing microarray gene expression data. In this module, upload the datasets from the online microarray database. A repository that contain the microarray gene expression data is then implement pre-processing steps to eliminate the irrelevant symbols and send for features selection.

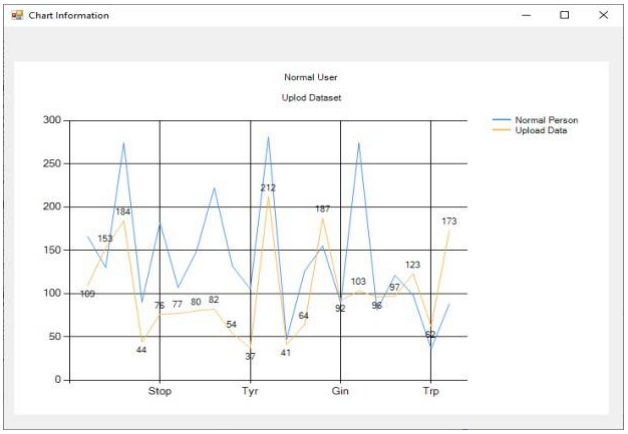
2.MEDIAN ESTIMATION

Median estimation is a common statistical method used to analyze data. Unlike the mean, the median is less influenced by extremely large or small values, making it a better representation of typical values. By declaring a predefined template as CTAG and calculating the frequency count for each symbol, median estimation can be performed.

3.FEATURE SELECTION

Feature selection can be done after uploading the datasets into the system and Implement PSO algorithm to splt the datasets as 2^2 combinations and perform median estimation based on predefined format as CTAG. And calculate the gene combination based on genetic code.





4.DISEASE PREDICTION

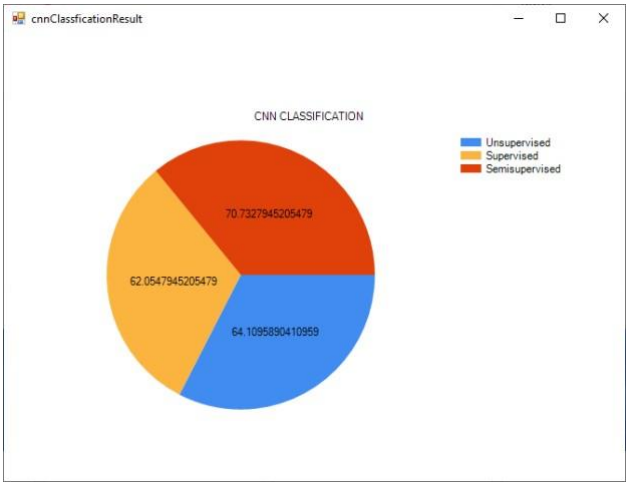
This module utilizes the Convolutional Neural Network (CNN) algorithm to classify various types of diseases based on gene expression. The CNN classifier has been known to perform exceptionally well in pattern recognition tasks over the recent years. The algorithm maps the input image into high dimensional feature space, constructs a hyperplane to exploit the margin of separation between classes, and identifies support vectors closest to the decision surface. If the classes are non-separable, the optimal hyperplane is chosen to minimize classification errors. The input image is first transformed into feature vectors and then mapped into the feature space using a kernel function. The CNN then computes the division in the feature space to separate the classes in the training data, and a global hyperplane is required to avoid overfitting. The CNN classifier can accurately differentiate between arteries and veins in the vessels.

5.SEVEAIRTY ANALYSIS

Severity analysis is done using a multiclass classification algorithm to determine the severity level of diseases. If the count exceeds a threshold, the severity is considered high, and patients are prescribed medication accordingly.

VII. RESULT AND DISCUSSION

CNN algorithm can be implemented and calculate the performance metrics for accuracy based on True positive rate, False positive rate, True negative rate and False negative rate. Accuracy rate is calculated as (TP+TN) / (TP+TN+FP+FN) and compare the results with existing unsupervised, supervised algorithms. The proposed semi-supervised algorithm provides improved accuracy rate than the existing algorithms. Then predicted the diseases with severity levels**.**



The performance result is shown, and CNN algorithm provides 70% accuracy.

IX. CONCLUSION AND FEATURE ENHANCEMENT

The method was designed to address the importance of gene ranking and selection prior to classification, which improves the prediction strength of the classifier. The project focused on promising accuracy results with very few numbers of gene subsets enabling the doctors to predict the type of cancer. The results on various disease datasets shows the importance of the same classifier used for both the gene selection and classification can improve the strength of the model. Then provide severity level for each classified disease. proposed a hybrid gene selection method, which combines a PSO methods and CNN classification to achieve high classification performance.

Future work includes partitioning of the original gene set into some distinct subsets or clusters so that the genes within a cluster are tightly coupled with strong association to the sample categories. We can extend the work to implement various classification algorithms to improve the accuracy rate at the time of disease prediction**.**

X. REFERNENCES

[1] Marcos-Zambrano, Laura Judith, et al. "Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment." Frontiers in microbiology (2021): 313.
[2] Soumya, Zayrit, et al. "The detection of Parkinson disease using the genetic algorithm and SVM classifier." Applied Acoustics 171 (2021): 107528.
[3] Zhao, Yinuan, Jia Cheng Huang, and Jianzhi Chen. "The integration of differentially expressed genes based on multiple microarray datasets for prediction of the prognosis in oral squamous cell carcinoma." Bioengineered 12.1 (2021): 3309-3321.
[4] Zhang, Lin, et al. "Identification of useful genes from multiple microarrays for ulcerative colitis diagnosis based on machine learning methods." Scientific reports 12.1 (2022): 1-13.
[5] Nasir, Muhammad Umar, et al. "Single and mitochondrial gene inheritance disorder prediction using machine learning." (2022).
[6] Thakur, Tanima, et al. "Gene expression-assisted cancer prediction techniques." Journal of Healthcare Engineering 2021 (2021).
[7] Karthik, S., and M. Sudha. "Predicting bipolar disorder and schizophrenia based on non-overlapping genetic phenotypes using deep neural network." Evolutionary Intelligence 14.2 (2021): 619-634.
[8] Gumaei, Abdu, et al. "Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression." Health Informatics Journal 27.1 (2021): 1460458221989402.
[9] Shu, Juan, et al. "Disease gene prediction with privileged information and heteroscedastic dropout." Bioinformatics 37. Supplement_1 (2021): i410-i417.
[10] Avsec-Žiga, , et al. "Effective gene expression prediction from sequence by integrating long-range interactions." Nature methods 18.10 (2021): 1196-1203.