IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10474

Gene Based Disease Prediction Using Deep Learning Techniques

M. Ananda Kumar[1] M.E., MBA., A. Manju[2] and S. DilipKumar[3] M.Tech., (Ph.D)

Assistant Professor, Department of Computer Science & Engineering, Arasu Engineering College, Kumbakonam, India[1]

PG Scholar, Department of Computer Science & Engineering, Arasu Engineering College, Kumbakonam, India [2]

Assistant Professor, Department of Computer Science & Engineering, Arasu Engineering College, Kumbakonam, India[3]

Abstract: Medical era has developed with a terrible level of achievements in disease pattern prediction, prevention and cure with the advancements of data mining techniques. Among various mining techniques feature selection occupies an indispensable role for the sake of improving accuracy in any kind of forecasting or cure of diseases. Thus it is treated as an essential earlier work of any kind of mining techniques. Microarray data have a high dimension of variables and a small sample size. In microarray data analyses, two important issues are how to choose genes, which provide reliable and good prediction for disease status, and how to determine the final gene set that is best for classification. Associations among genetic markers mean one can exploit information redundancy to potentially reduce classification cost in terms of time and computation cost. So in this project, we can implement the framework to predict the diseases using optimization and classification algorithm such as Genetic algorithm and Semi-supervised deep learning algorithm with improved accuracy rate. This paper presents an overview of various disease classification methods and evaluates these proposed methods based on their classification accuracy, computational time and ability to reveal gene information. We have also evaluated and introduced various proposed gene selection method.

Keywords: Microarray data, Bio-markers. High Dimension data, Optimization, Deep learning algorithm

I. INTRODUCTION

Microarray technology has become one of the indispensable tools that many biologists use to monitor genome wide expression levels of genes in a given organism. A microarray is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations called spots (or features). A microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. The DNA in a spot may either be genomic DNA or short stretch of oligo-nucleotide strands that correspond to a gene. The spots are printed on to the glass slide by a robot or are synthesized by the process of photolithography. Microarrays may be used to measure gene expression in many ways, but one of the most popular applications is to compare expression of a set of genes from a cell maintained in a particular condition (condition A) to the same set of genes from a reference cell maintained under normal conditions (condition B). Clustering techniques have proven to be helpful to understand gene function, gene regulation, cellular processes, and subtypes of cells. Genes with similar expression patterns (co-expressed genes) can be clustered together with similar cellular functions. This approach may further understanding of the functions of many genes for which information has not been previously available. Furthermore, co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation of expression patterns between those genes indicates co-regulation. Searching for common DNA sequences at the promoter regions of genes within the same cluster allows regulatory motifs specific to each gene cluster to be identified and cis-regulatory elements to be proposed. The inference of regulation through the clustering of gene expression data also gives rise to hypotheses regarding the mechanism of the transcriptional regulatory network. Finally, clustering different samples on the basis of corresponding expression profiles may reveal sub-cell types which are hard to identify by traditional morphology-based approaches.

Due to the special characteristics of gene expression data, and the particular requirements from the biological domain, gene-based clustering presents several new challenges and is still an open problem. First, cluster analysis is typically the first step in data mining and knowledge discovery. The purpose of clustering gene expression data is to reveal the natural data structures and gain some initial insights regarding data distribution. Therefore, a good clustering algorithm should depend as little as possible on prior knowledge, which is usually not available before cluster analysis. One requiring the pre-determined number of clusters. Second, due to the complex procedures of microarray experiments, gene expression data often contain a huge amount of noise. The basic structure gene symbols are shown in figure 1.

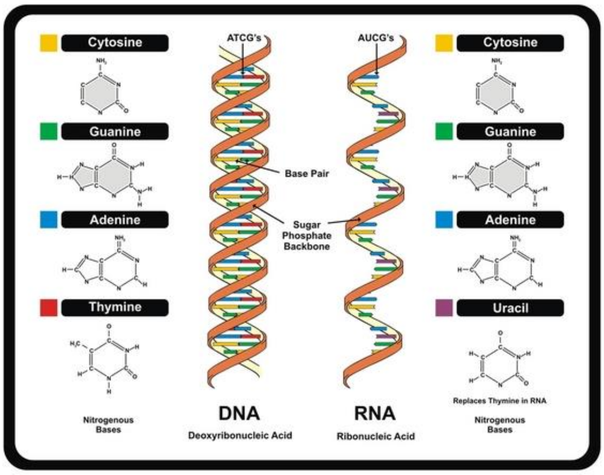


Fig. 1 Gene Symbols

II. Related Work

“A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets." deals with the problem of feature subset selection (FSS) in the task of supervised classification. Supervised classification is probably the data-mining/machine-learning technique that is most commonly used in practice [1]. Feature subset selection is a key problem in the data-mining classification task that helps to obtain more compact and understandable models without degrading (or even improving) their performance. In this work we focus on FSS in high-dimensional datasets, that is, with a very large number of predictive attributes. In this case, standard sophisticated wrapper algorithms cannot be applied because of their complexity, and computationally lighter filter-wrapper algorithms have recently been proposed. In this work we propose a stochastic algorithm based on the GRASP meta-heuristic, with the main goal of speeding up the feature subset selection process, basically by reducing the number of wrapper evaluations to carry out. GRASP is a multi-start constructive method which constructs a solution in its first stage, and then runs an improving stage over that solution. GRASP can then be viewed as a global optimization method, which usually improves the quality of the solution obtained by specific heuristics.

“Time series prediction using RBF neural networks with a nonlinear time-varying evolution PSO algorithm” analyzed the time series data which is an ordered sequence of observations that are evenly spaced at uniform time intervals and measured successively. Prediction of time series uses a sequence of historical values to develop a model for forecasting future values[2]. The quality of load forecasts has significant impact on the economic operation of the electric utility since many decisions are based on these forecasts. Accurate prediction will reduce the risk of management decision making in government and industry. Recently, time series prediction has received considerable applications in various filed, such as economics and business, biology, medicine, chemical, and engineering. Load forecast has been widely concerned in modern power system planning, operation, and control. According to the lead time of forecast, load forecast is divided into short-term, middle- term, and long-term load forecast. Generally, load forecast is trended to short-term prediction such as one-day ahead prediction, since longer period prediction (middle-term or long-term) may not be reliant due to error propagation. The time series prediction of a practical power system is investigated in this paper. The radial basis function neural network (RBFNN) with a nonlinear time-varying evolution particle swarm optimization (NTVE-PSO) algorithm is developed.

“Evolving artificial neural networks using an improved PSO and DPSO" developed the advent of evolutionary algorithms (EAs) has inspired new resources for optimization problem solving, such as the optimal design of artificial neural networks (ANNs) and fuzzy systems. Since EAs are heuristic and stochastic based on populations made up of individuals with a specified behavior similar to biological phenomenon, they are robust and efficient at exploring an entire solution space of optimization problems. EAs have been successfully used to evolve weights, structure, and learning parameters of ANNs in recent years. Yao and Liu proposed a prominent evolutionary ANN approach called EPNet by using evolutionary programming (EP) algorithm [3]. Weights and structure are evolved simultaneously by using partial training, mutation of weights and addition or removal of connections or nodes. EPNet

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10474

encourages smaller networks, as removals are attempted before additions, and behavioral link is maintained between parents and offspring through partial training and node splitting. Castillo et al. proposed a method (G-Prop, genetic back-propagation (BP)) that attempts to search the initial weights and hidden-layer size of multilayer perceptrons (MLPs). The application of the G-Prop algorithm to several real-world and benchmark problems showed that MLPs evolved by G-Prop are smaller and achieve a better generalization than other perceptron training algorithms. Palmes et al. proposed a mutation-based genetic ANN (MGNN) algorithm.

“A modified binary particle swarm optimization for knapsack problems." implemented the particle swarm optimization algorithm, originally introduced in terms of social and cognitive behavior by Kennedy and Eberhart, solves problems in many fields, especially engineering and computer science. Only within a few years of its introduction PSO has gained wide popularity as a powerful global optimization tool and is competing with well-established population based search algorithms [4]. The inspiration behind the development of PSO is the mechanism by which the birds in a flock and the fishes in a school cooperate while searching for food. In PSO, a group of active, dynamic and interactive members called swarm produces a very intelligent search behaviour using collaborative trial and error. Each member of the swarm called particle, represents a potential solution of the problem under consideration. Each particle in the swarm relies on its own experience as well as the experience of its best neighbour (in terms of fitness). Each particle has an associated fitness value. These particles move through search space with a specified velocity in search of optimal solution. Each particle maintains a memory which helps it in keeping the track of the best position it has achieved so far. This is called the particle’s personal best position (pbest) and the best position the swarm has achieved so far is called global best position (gbest). The movement of the particles is influenced by two factors using information from iteration-to-iteration as well as particle-to-particle.

“An agent-based clustering approach for gene selection in gene expression microarray" analyzed gene selection (GS) approach which can be generically defined as the process of extracting gene subsets whose expression level values are representative of a particular target feature, i.e., clinical or biological annotation. GS is a very active research area in the analysis of gene expression microarray, which is contributing to the development of the field as a result of involved data mining and machine learning techniques. Particularly [5], GS from microarrays is addressed to identify/discover those genes which are expressed differentially according to a determined target disease (namely informative genes). GS methods have been divided into the following four categories: filters, wrappers, embedded and ensemble. Filter methods have been directed to discriminate or filter features/ genes based on the intrinsic properties of the dataset. They do this by estimating their relevance scores to state a cut-off schema where an upper/lower bound is imposed to choose features with the best scores. Wrapper methods use a classifier to find the most discriminant feature subset by minimizing an error prediction function. Embedded methods are similar to wrapper but additionally they interact with the learning model, which reduces the runtime taken by wrapper methods.

III. PROBLEM DEFINITION

Cancer research is one of the major research areas in the medical field. Accurate prediction of different tumor types has great value in providing better treatment and toxicity minimization on the patients. Different classification methods from statistical and machine learning area have been applied to cancer classification, but there are some issues that make it a nontrivial task. The gene expression data is very different from any of the data these methods had previously dealt with. First, it has very high dimensionality, usually contains thousands to tens of thousands of genes. Second, publicly available data size is very small, all below 100. Third, most genes are irrelevant to cancer distinction. It is obvious that those existing classification methods were not designed to handle this kind of data efficiently and effectively. Some researchers proposed to do gene selection prior to cancer classification. Performing gene selection helps to reduce data size thus improving the running time. In this existing system, we present a comprehensive overview of various cancer classification methods and evaluate them based on their computation time, classification accuracy and ability to reveal biologically meaningful gene information. We also introduce and evaluate various gene selection methods which we believe should be an integral preprocessing step for cancer classification. In order to obtain a full picture of cancer classification, we also discuss several issues related to cancer classification, including the biological significance vs. statistical significance of a cancer classifier, the asymmetrical classification errors for cancer classifiers, and the gene contamination problem.

IV. IMPLEMENTATION WORK

Microarray technology has made the modern biological research by permitting the simultaneous study of genes comprising a large part of the genome. 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. Microarrays are capable of determining the expression levels of thousands of genes

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10474

simultaneously. One important application of gene expression data is classification of samples into categories. In combination with classification methods, this tool can be useful to support clinical management decisions for individual patients, e.g. in oncology. Standard statistic methodologies in classification or prediction do not work well when the number of variables p (genes) far too exceeds the number of samples n which is the case in gene microarray expression data. The goal of our proposed project will be to use supervised learning to classify and predict diseases, based on the gene expressions collected from microarrays. Known sets of data will be used to train the deep learning protocols to categorize diseases according to their gene patterns. The outcome of this study will provide information regarding the efficiency of the machine learning techniques, in particular Convolutional Neural network method. The efficiency of classification depends on the type of kernel function that is used. So here we will analyses the performance of various kernel functions used for classification purpose. Finally predict the diseases with severity levels and predict various types of diseases. The proposed layout is shown in figure 2.

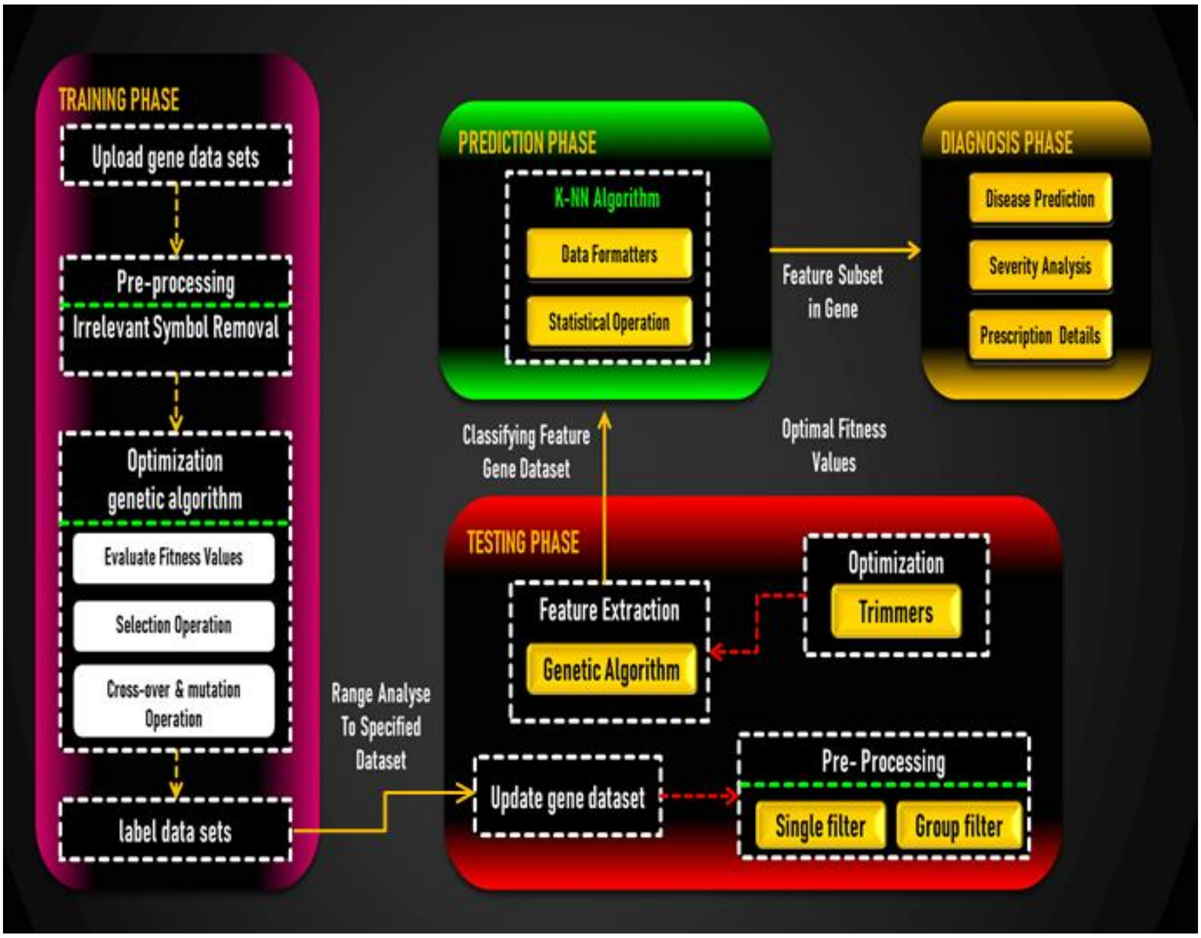


Fig. 2 System architecture

This figure 2 gives the overall architecture diagram of proposed work. In this framework include training and testing phase. Training phase we can upload the gene data and using preprocessing steps to eliminate the irrelevant symbols. Then extract the features using Genetic algorithm to select the fitness values. And in testing phase, we can input the gene datasets and perform preprocessing and features extraction steps. Finally the classify the diseases using Convolutional neural network algorithm with improved accuracy.

A. *Datasets acquisition*

A microarray database is a repository containing microarray gene expression data. The key uses of a microarray database are to store the measurement data, manage a searchable index, and make the data available to other applications for analysis and interpretation. In this module, upload the datasets. The dataset may be microarray dataset. A microarray database is a repository containing microarray gene expression data. Then implement pre-processing steps to eliminate the irrelevant symbols.

B. *Preprocessing*

The median is the value separating the higher half from the lower half of a data sample (a population or a probability distribution). For a data set, it may be thought of as the "middle" value. The median is a commonly used measure of the properties of a data set in statistics and probability theory. The basic advantage of the median in describing data compared to the mean (often simply described as the "average") is that it is not skewed so much by a small proportion of extremely large or small values, and so it may give a better idea of a "typical" value. In this module split the gene symbols into $2^{\wedge 2}$combinations. And eliminate the header symbol for future calculation. Declare the predefined Template as CTAG and calculate frequency count for each symbol.

C. *Optimization*

In Genetic algorithm, can analyses coverage of the data before clustering begins. And propose an algorithm, which modifies the nearest centroid sorting and the transfer algorithm, of the spatial medians clustering. It has two distinct phases: one of transferring an object from one cluster to another and the other of amalgamating the single member cluster with it's the nearest cluster. Given a starting partition, each possible transfer is tested in turn to see if it would improve the value of clustering criterion. When no further transfers can improve the criterion value, each possible amalgamation of the single member cluster and other clusters is tested.

D. *Classification*

Classifiers based on gene expression are generally probabilistic, that is they only predict that a certain percentage of the individuals that have a given expression profile will also have the phenotype, or outcome, of interest. Therefore, statistical validation is necessary before models can be employed, especially in clinical settings. In this module implement convolutional neural network algorithm to classify the various types of diseases from gene expression. Classification is done with the help of K-NN classifier. In the recent years, K-NN classifiers have established excellent performance in a variety of pattern recognition troubles. The input space is planned into a high dimensional feature space. Then, the hyper plane that exploits the margin of separation between classes is constructed. The points that lie closest to the decision surface are called support vectors directly involves its location. When the classes are non-separable, the optimal hyper plane is the one that minimizes the probability of classification error. Initially input image is formulated in feature vectors. Then these feature vectors mapped with the help of kernel function in the feature space. And finally division is computed in the feature space to separate out the classes for training data. A global hyper plane is required by the K-NN in order to divide both the program of examples in training set and avoid over fitting. This phenomenon of K-NN is higher in comparison to other machine learning techniques which are based on artificial intelligence. Here the important feature for the classification is the width of the vessels. With the help of K-NN classifier we can easily separate out the vessels into arteries and veins. The K-NNs demonstrate various attractive features such as good generalization ability compared to other classifiers. Indeed, there are relatively few free parameters to adjust and it is not required to find the architecture experimentally. The K-NNs algorithm separates the classes of input patterns with the maximal margin hyper plane. This hyper plane is constructed as:

$$f(x) = \langle w, x \rangle + b$$

To benefit from non-linear decision boundaries the separation is performed in a feature space F, which is introduced by a nonlinear mapping φ the input patterns. This mapping is defined as follows:

$$\langle \varphi\ (x_1), \varphi(x_2) \rangle = K(x_1, x_2)\ \forall (x_1, x_2) \in X$$

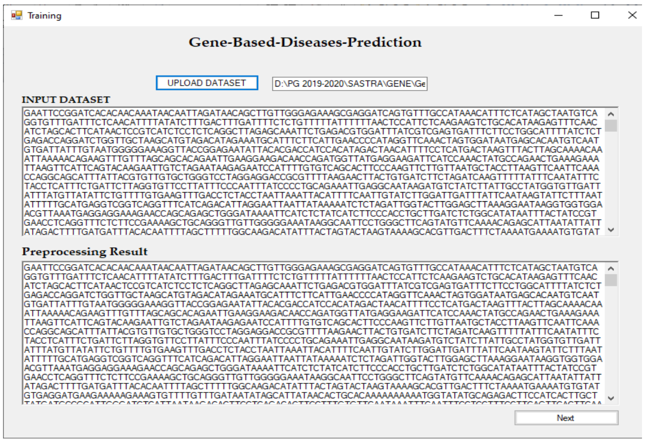
for some kernel function K (·, ·). The kernel function represents the non-linear transformation of the original feature space into the F.

E. *Disease prediction & prescription*

Using multi class classification algorithm to classify the severity level of diseases using classified data count. If count is more than threshold means, provide severity as high and count is less than threshold means, consider as normal. Then provide prescription to patients according to the diseases. And also evaluate the performance of the system in terms of accuracy rate.

V. EXPERIMENTAL RESULTS

We can implement the system in C#.NET framework and SQL SERVER as back end. The Gene dataset upload as text files. It contains various symbols exampled as "ATCG". The following figures are shown the layout of the proposed work.



Training

Gene-Based-Diseases-Prediction

UPLOAD DATASET D:\PG 2019-2020\SASTRA\GENE\Ge

INPUT DATASET

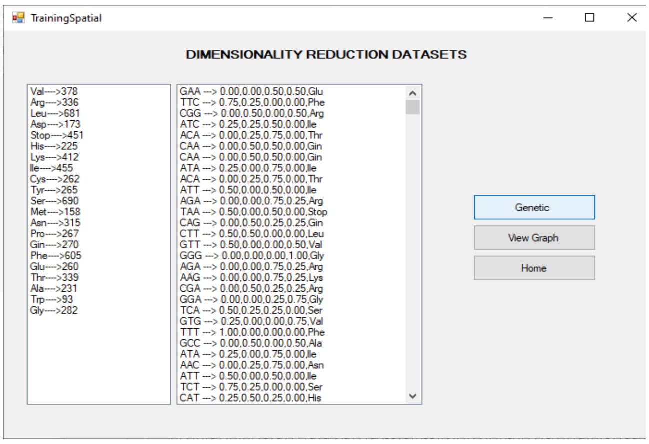
```
GAATTCCGGATCACACAACAAATAACAATTAGATAACAGCTTGTTGGGAGAAAGCGAGGATCAGTGTTTGCCATAAACATTTCTCATAGCTAATGTCA
GGTGTTTGATTTCTCAACATTTTATATCTTTGACTTTGATTTTCTCTGTTTTTATTTTTTAACTCCATTCTCAAGAAGTCTGCACATAAGAGTTTCAAC
ATCTAGCACTTCATAACTCCGTCATCTCCTCTCAGGCTTAGAGCAAATTCTGAGACGTGGATTTATCGTCGAGTGATTTCTTCCTGGCATTTTATCTCT
GAGACCAGGATCTGGTTGCTAAGCATGTAGACATAGAAATGCATTTCTTCATTGAACCCCATAGGTTCAAACTAGTGGATAATGAGCACAATGTCAAT
GTGATTATTTGTAATGGGGGAAAGGTTACCGGAGAATATTACACGACCATCCACATAGACTAACATTTTCCTCATGACTAAGTTTACTTAGCAAAACAA
ATTAAAAACAGAAGTTTGTTTAGCAGCACAGAATTGAAGGAAGACAACCAGATGGTTATGAGGAAGATTCATCCAAACTATGCCAGAACTGAAAGAAA
TTAAGTTCATTCAGTACAAGAATTGTCTAGAATAAGAGAATCCATTTTGTGTCAGCACTTCCCAAGTTCTTGTTAATGCTACCTTAAGTTCAATTCAAA
CCAGGCAGCATTTATTACGTGTTGTGCTGGGTCCTAGGAGGACCGCGTTTTAAGAACTTACTGTGATCTTCTAGATCAAGTTTTTATTTCAATATTTC
TACCTCATTTCTGATTCTTAGGTGTTCCTTATTTCCCAATTTATCCCCTGCAGAAATTGAGGCAATAAGATGTCTATCTTATTGCCTATGGTGTTGATT
ATTTATGTTATATTCTGTTTTGTGAAGTTTGACCTCTACCTAATTAAATTACATTTTCAATTGTATCTTGGATTGATTTATTCAATAAGTATTCTTTAAT
ATTTTTGCATGAGGTCGGTCAGGTTTCATCAGACATTAGGAATTAATTATAAAAATCTCTAGATTGGTACTTGGAGCTTAAAGGAATAAGGTGGTGGA
ACGTTAAATGAGGAGGAAAGAACCAGCAGAGCTGGGATAAAATTCATCTCTATCATCTTCCCACCTGCTTGATCTCTGGCATATAATTTACTATCCGT
GAACCTCAGGTTTCTCTTCCGAAAAGCTGCAGGGTTGTTGGGGGAAATAAGGCAATTCCTGGGCTTCAGTATGTTCAAAACAGAGCATTAATATTATT
ATAGACTTTTGATGATTTACACAATTTTAGCTTTTTGGCAAGACATATTTACTAGTACTAAGTAAAAGCACGTTGACTTTCTAAAATGAAAATGTGTAT
```

Preprocessing Result

```
GAATTCCGGATCACACAACAAATAACAATTAGATAACAGCTTGTTGGGAGAAAGCGAGGATCAGTGTTTGCCATAAACATTTCTCATAGCTAATGTCA
GGTGTTTGATTTCTCAACATTTTATATCTTTGACTTTGATTTTCTCTGTTTTTATTTTTTAACTCCATTCTCAAGAAGTCTGCACATAAGAGTTTCAAC
ATCTAGCACTTCATAACTCCGTCATCTCCTCTCAGGCTTAGAGCAAATTCTGAGACGTGGATTTATCGTCGAGTGATTTCTTCCTGGCATTTTATCTCT
GAGACCAGGATCTGGTTGCTAAGCATGTAGACATAGAAATGCATTTCTTCATTGAACCCCATAGGTTCAAACTAGTGGATAATGAGCACAATGTCAAT
GTGATTATTTGTAATGGGGGAAAGGTTACCGGAGAATATTACACGACCATCCACATAGACTAACATTTTCCTCATGACTAAGTTTACTTAGCAAAACAA
ATTAAAAACAGAAGTTTGTTTAGCAGCACAGAATTGAAGGAAGACAACCAGATGGTTATGAGGAAGATTCATCCAAACTATGCCAGAACTGAAAGAAA
TTAAGTTCATTCAGTACAAGAATTGTCTAGAATAAGAGAATCCATTTTGTGTCAGCACTTCCCAAGTTCTTGTTAATGCTACCTTAAGTTCAATTCAAA
CCAGGCAGCATTTATTACGTGTTGTGCTGGGTCCTAGGAGGACCGCGTTTTAAGAACTTACTGTGATCTTCTAGATCAAGTTTTTATTTCAATATTTC
TACCTCATTTCTGATTCTTAGGTGTTCCTTATTTCCCAATTTATCCCCTGCAGAAATTGAGGCAATAAGATGTCTATCTTATTGCCTATGGTGTTGATT
ATTTATGTTATATTCTGTTTTGTGAAGTTTGACCTCTACCTAATTAAATTACATTTTCAATTGTATCTTGGATTGATTTATTCAATAAGTATTCTTTAAT
ATTTTTGCATGAGGTCGGTCAGGTTTCATCAGACATTAGGAATTAATTATAAAAATCTCTAGATTGGTACTTGGAGCTTAAAGGAATAAGGTGGTGGA
ACGTTAAATGAGGAGGAAAGAACCAGCAGAGCTGGGATAAAATTCATCTCTATCATCTTCCCACCTGCTTGATCTCTGGCATATAATTTACTATCCGT
GAACCTCAGGTTTCTCTTCCGAAAAGCTGCAGGGTTGTTGGGGGAAATAAGGCAATTCCTGGGCTTCAGTATGTTCAAAACAGAGCATTAATATTATT
ATAGACTTTTGATGATTTACACAATTTTAGCTTTTTGGCAAGACATATTTACTAGTACTAAGTAAAAGCACGTTGACTTTCTAAAATGAAAATGTGTAT
GTGAGGATGAAGAAAAAGAAAGTGTTTTGTTTGATAATATAGCATTATAACACTGCACAAAAAAAAAATGGTATATGCAGAGACTTCCATCACTTGCT
```

Next

Fig 3 Data with preprocessing



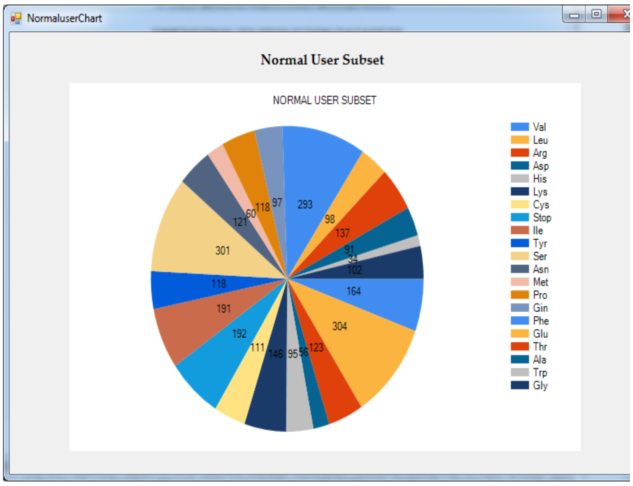
TrainingSpatial

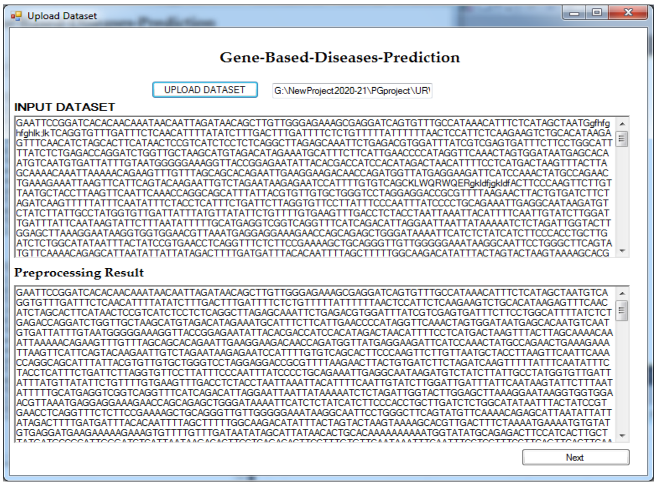
DIMENSIONALITY REDUCTION DATASETS

```
Val---->378
Arg---->336
Leu---->681
Asp---->173
Stop---->451
His---->225
Lys---->412
Ile---->455
Cys---->262
Tyr---->265
Ser---->690
Met---->158
Asn---->315
Pro---->267
Gin---->270
Phe---->605
Glu---->260
Thr---->339
Ala---->231
Trp---->93
Gly---->282
```

```
GAA ---> 0.00,0.00,0.50,0.50,Glu
TTC ---> 0.75,0.25,0.00,0.00,Phe
CGG ---> 0.00,0.50,0.00,0.50,Arg
ATC ---> 0.25,0.25,0.50,0.00,Ile
ACA ---> 0.00,0.25,0.75,0.00,Thr
CAA ---> 0.00,0.50,0.50,0.00,Gin
CAA ---> 0.00,0.50,0.50,0.00,Gin
ATA ---> 0.25,0.00,0.75,0.00,Ile
ACA ---> 0.00,0.25,0.75,0.00,Thr
ATT ---> 0.50,0.00,0.50,0.00,Ile
AGA ---> 0.00,0.00,0.75,0.25,Arg
TAA ---> 0.50,0.00,0.50,0.00,Stop
CAG ---> 0.00,0.50,0.25,0.25,Gin
CTT ---> 0.50,0.50,0.00,0.00,Leu
GTT ---> 0.50,0.00,0.00,0.50,Val
GGG ---> 0.00,0.00,0.00,1.00,Gly
AGA ---> 0.00,0.00,0.75,0.25,Arg
AAG ---> 0.00,0.00,0.75,0.25,Lys
CGA ---> 0.00,0.50,0.25,0.25,Arg
GGA ---> 0.00,0.00,0.25,0.75,Gly
TCA ---> 0.50,0.25,0.25,0.00,Ser
GTG ---> 0.25,0.00,0.00,0.75,Val
TTT ---> 1.00,0.00,0.00,0.00,Phe
GCC ---> 0.00,0.50,0.00,0.50,Ala
ATA ---> 0.25,0.00,0.75,0.00,Ile
AAC ---> 0.00,0.25,0.75,0.00,Asn
ATT ---> 0.50,0.00,0.50,0.00,Ile
TCT ---> 0.75,0.25,0.00,0.00,Ser
CAT ---> 0.25,0.50,0.25,0.00,His
```

Fig 4 Dimensionality reduction





Upload Dataset

Gene-Based-Diseases-Prediction

UPLOAD DATASET G:\NewProject2020-21\PGproject\UR\

INPUT DATASET

```
GAATTCCGGATCACACAACAAATAACAATTAGATAACAGCTTGTTGGGAGAAAGCGAGGATCAGTGTTTGCCATAAACATTTCTCATAGCTAATGgfhfg
hfghlk;lkTCAGGTGTTTGATTTCTCAACATTTTATATCTTTGACTTTGATTTTCTCTGTTTTTATTTTTTAACTCCATTCTCAAGAAGTCTGCACATAAGA
GTTTCAACATCTAGCACTTCATAACTCCGTCATCTCCTCTCAGGCTTAGAGCAAATTCTGAGACGTGGATTTATCGTCGAGTGATTTCTTCCTGGCATT
TTATCTCTGAGACCAGGATCTGGTTGCTAAGCATGTAGACATAGAAATGCATTTCTTCATTGAACCCCATAGGTTCAAACTAGTGGATAATGAGCACA
ATGTCAATGTGATTATTTGTAATGGGGGAAAGGTTACCGGAGAATATTACACGACCATCCACATAGACTAACATTTTCCTCATGACTAAGTTTACTTA
GCAAAACAAATTAAAAACAGAAGTTTGTTTAGCAGCACAGAATTGAAGGAAGACAACCAGATGGTTATGAGGAAGATTCATCCAAACTATGCCAGAAC
TGAAAGAAATTAAGTTCATTCAGTACAAGAATTGTCTAGAATAAGAGAATCCATTTTGTGTCAGCKLWQRWQERgkldfjgkldfACTTCCCAAGTTCTTGT
TAATGCTACCTTAAGTTCAATTCAAACCAGGCAGCATTTATTACGTGTTGTGCTGGGTCCTAGGAGGACCGCGTTTTAAGAACTTACTGTGATCTTCT
AGATCAAGTTTTTATTTCAATATTTCTACCTCATTTCTGATTCTTAGGTGTTCCTTATTTCCCAATTTATCCCCTGCAGAAATTGAGGCAATAAGATGT
CTATCTTATTGCCTATGGTGTTGATTATTTATGTTATATTCTGTTTTGTGAAGTTTGACCTCTACCTAATTAAATTACATTTTCAATTGTATCTTGGAT
TGATTTATTCAATAAGTATTCTTTAATATTTTTGCATGAGGTCGGTCAGGTTTCATCAGACATTAGGAATTAATTATAAAAATCTCTAGATTGGTACTT
GGAGCTTAAAGGAATAAGGTGGTGGAACGTTAAATGAGGAGGAAAGAACCAGCAGAGCTGGGATAAAATTCATCTCTATCATCTTCCCACCTGCTTG
ATCTCTGGCATATAATTTACTATCCGTGAACCTCAGGTTTCTCTTCCGAAAAGCTGCAGGGTTGTTGGGGGAAATAAGGCAATTCCTGGGCTTCAGTA
TGTTCAAAACAGAGCATTAATATTATTATAGACTTTTGATGATTTACACAATTTTAGCTTTTTGGCAAGACATATTTACTAGTACTAAGTAAAAGCACG
```

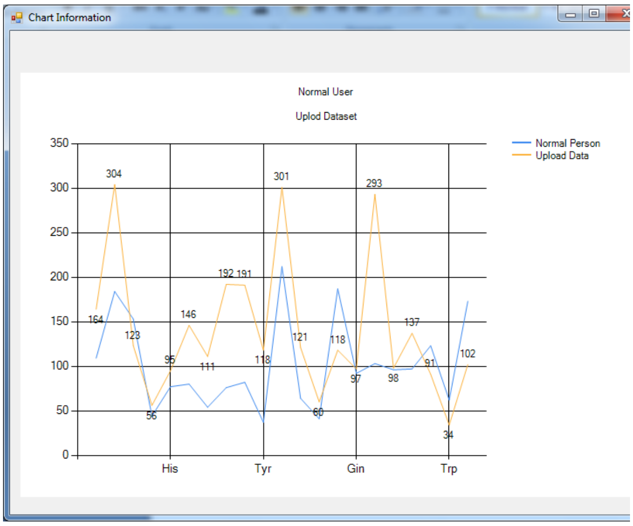
Preprocessing Result

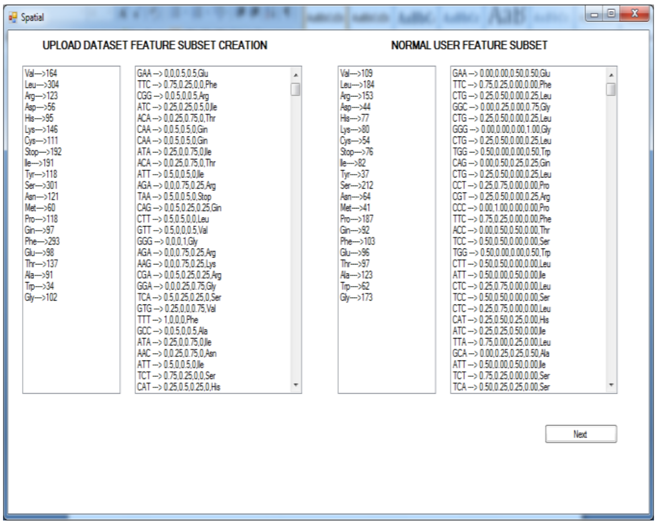
```
GAATTCCGGATCACACAACAAATAACAATTAGATAACAGCTTGTTGGGAGAAAGCGAGGATCAGTGTTTGCCATAAACATTTCTCATAGCTAATGTCA
GGTGTTTGATTTCTCAACATTTTATATCTTTGACTTTGATTTTCTCTGTTTTTATTTTTTAACTCCATTCTCAAGAAGTCTGCACATAAGAGTTTCAAC
ATCTAGCACTTCATAACTCCGTCATCTCCTCTCAGGCTTAGAGCAAATTCTGAGACGTGGATTTATCGTCGAGTGATTTCTTCCTGGCATTTTATCTCT
GAGACCAGGATCTGGTTGCTAAGCATGTAGACATAGAAATGCATTTCTTCATTGAACCCCATAGGTTCAAACTAGTGGATAATGAGCACAATGTCAAT
GTGATTATTTGTAATGGGGGAAAGGTTACCGGAGAATATTACACGACCATCCACATAGACTAACATTTTCCTCATGACTAAGTTTACTTAGCAAAACAA
ATTAAAAACAGAAGTTTGTTTAGCAGCACAGAATTGAAGGAAGACAACCAGATGGTTATGAGGAAGATTCATCCAAACTATGCCAGAACTGAAAGAAA
TTAAGTTCATTCAGTACAAGAATTGTCTAGAATAAGAGAATCCATTTTGTGTCAGCACTTCCCAAGTTCTTGTTAATGCTACCTTAAGTTCAATTCAAA
CCAGGCAGCATTTATTACGTGTTGTGCTGGGTCCTAGGAGGACCGCGTTTTAAGAACTTACTGTGATCTTCTAGATCAAGTTTTTATTTCAATATTTC
TACCTCATTTCTGATTCTTAGGTGTTCCTTATTTCCCAATTTATCCCCTGCAGAAATTGAGGCAATAAGATGTCTATCTTATTGCCTATGGTGTTGATT
ATTTATGTTATATTCTGTTTTGTGAAGTTTGACCTCTACCTAATTAAATTACATTTTCAATTGTATCTTGGATTGATTTATTCAATAAGTATTCTTTAAT
ATTTTTGCATGAGGTCGGTCAGGTTTCATCAGACATTAGGAATTAATTATAAAAATCTCTAGATTGGTACTTGGAGCTTAAAGGAATAAGGTGGTGGA
ACGTTAAATGAGGAGGAAAGAACCAGCAGAGCTGGGATAAAATTCATCTCTATCATCTTCCCACCTGCTTGATCTCTGGCATATAATTTACTATCCGT
GAACCTCAGGTTTCTCTTCCGAAAAGCTGCAGGGTTGTTGGGGGAAATAAGGCAATTCCTGGGCTTCAGTATGTTCAAAACAGAGCATTAATATTATT
ATAGACTTTTGATGATTTACACAATTTTAGCTTTTTGGCAAGACATATTTACTAGTACTAAGTAAAAGCACGTTGACTTTCTAAAATGAAAATGTGTAT
GTGAGGATGAAGAAAAAGAAAGTGTTTTGTTTGATAATATAGCATTATAACACTGCACAAAAAAAAAATGGTATATGCAGAGACTTCCATCACTTGCT
```

Next

Fig. 5 Normal user subset

Fig. 6 Testing dataset





Spatial

UPLOAD DATASET FEATURE SUBSET CREATION

```
Val--->164
Leu--->304
Arg--->123
Asp--->56
His--->95
Lys--->146
Cys--->111
Stop--->192
Ile--->191
Tyr--->118
Ser--->301
Asn--->121
Met--->60
Pro--->118
Gin--->97
Phe--->293
Glu--->98
Thr--->137
Ala--->91
Trp--->34
Gly--->102
```

```
GAA --> 0,0,0.5,0.5,Glu
TTC --> 0.75,0.25,0,0,Phe
CGG --> 0,0.5,0,0.5,Arg
ATC --> 0.25,0.25,0.5,0,Ile
ACA --> 0,0.25,0.75,0,Thr
CAA --> 0,0.5,0.5,0,Gin
CAA --> 0,0.5,0.5,0,Gin
ATA --> 0.25,0,0.75,0,Ile
ACA --> 0,0.25,0.75,0,Thr
ATT --> 0.5,0,0.5,0,Ile
AGA --> 0,0,0.75,0.25,Arg
TAA --> 0.5,0,0.5,0,Stop
CAG --> 0,0.5,0.25,0.25,Gin
CTT --> 0.5,0.5,0,0,Leu
GTT --> 0.5,0,0,0.5,Val
GGG --> 0,0,0,1,Gly
AGA --> 0,0,0.75,0.25,Arg
AAG --> 0,0,0.75,0.25,Lys
CGA --> 0,0.5,0.25,0.25,Arg
GGA --> 0,0,0.25,0.75,Gly
TCA --> 0.5,0.25,0.25,0,Ser
GTG --> 0.25,0,0,0.75,Val
TTT --> 1,0,0,0,Phe
GCC --> 0,0.5,0,0.5,Ala
ATA --> 0.25,0,0.75,0,Ile
AAC --> 0,0.25,0.75,0,Asn
ATT --> 0.5,0,0.5,0,Ile
TCT --> 0.75,0.25,0,0,Ser
CAT --> 0.25,0.5,0.25,0,His
```

NORMAL USER FEATURE SUBSET

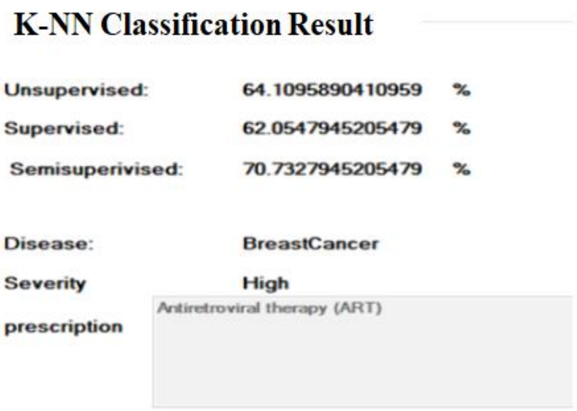
```
Val--->109
Leu--->184
Arg--->153
Asp--->44
His--->77
Lys--->80
Cys--->54
Stop--->76
Ile--->82
Tyr--->37
Ser--->212
Asn--->64
Met--->41
Pro--->187
Gin--->92
Phe--->103
Glu--->96
Thr--->97
Ala--->123
Trp--->62
Gly--->173
```

```
GAA --> 0.00,0.00,0.50,0.50,Glu
TTC --> 0.75,0.25,0.00,0.00,Phe
CTG --> 0.25,0.50,0.00,0.25,Leu
GGC --> 0.00,0.25,0.00,0.75,Gly
CTG --> 0.25,0.50,0.00,0.25,Leu
GGG --> 0.00,0.00,0.00,1.00,Gly
CTG --> 0.25,0.50,0.00,0.25,Leu
TGG --> 0.50,0.00,0.00,0.50,Trp
CAG --> 0.00,0.50,0.25,0.25,Gin
CTG --> 0.25,0.50,0.00,0.25,Leu
CCT --> 0.25,0.75,0.00,0.00,Pro
CGT --> 0.25,0.50,0.00,0.25,Arg
CCC --> 0.00,1.00,0.00,0.00,Pro
TTC --> 0.75,0.25,0.00,0.00,Phe
ACC --> 0.00,0.50,0.50,0.00,Thr
TCC --> 0.50,0.50,0.00,0.00,Ser
TGG --> 0.50,0.00,0.00,0.50,Trp
CTT --> 0.50,0.50,0.00,0.00,Leu
ATT --> 0.50,0.00,0.50,0.00,Ile
CTC --> 0.25,0.75,0.00,0.00,Leu
TCC --> 0.50,0.50,0.00,0.00,Ser
CTC --> 0.25,0.75,0.00,0.00,Leu
CAT --> 0.25,0.50,0.25,0.00,His
ATC --> 0.25,0.25,0.50,0.00,Ile
TTA --> 0.75,0.00,0.25,0.00,Leu
GCA --> 0.00,0.25,0.25,0.50,Ala
ATT --> 0.50,0.00,0.50,0.00,Ile
TCT --> 0.75,0.25,0.00,0.00,Ser
TCA --> 0.50,0.25,0.25,0.00,Ser
```

Next

Fig. 7 Graph for normal and uploaded data

Fig. 7: Classification



K-NN Classification Result

Unsupervised: 64.1095890410959 %

Supervised: 62.0547945205479 %

Semisuperivised: 70.7327945205479 %

Disease: BreastCancer

Severity High

prescription Antiretroviral therapy (ART)

Fig. 8 Disease prediction

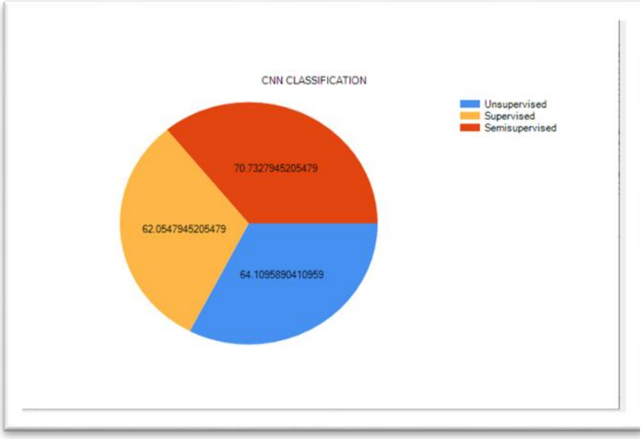


Fig. 9 Performance chart

VI. CONCLUSION

The median is the value separating the higher half from the lower half of a data sample. For a data set, it may be thought of as the "middle" value. The median is a commonly used measure of the properties of a data set in K-NN statistics and probability theory. The basic advantage of the median in describing data compared to the mean (often simply described as the "average") is that it is not skewed so much by a small proportion of extremely large or small values, and so it may give a better idea of a "typical" value. In this module split the gene symbols into 2^2 combinations. And eliminate the header symbol for future calculation. Declare the predefined Template as CTAG and calculate frequency count for each symbol.

REFERENCES

[1] Bermejo, Pablo, Jose A. Gámez, and Jose M. Puerta. "A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets." Pattern Recognition Letters 32.5 (2011): 701-711.

[2] Lee, Cheng-Ming, and Chia-Nan Ko. "Time series prediction using RBF neural networks with a nonlinear time-varying evolution PSO algorithm." Neurocomputing 73.1-3 (2009): 449-460.

[3] Yu, Jianbo, Shijin Wang, and Lifeng Xi. "Evolving artificial neural networks using an improved PSO and DPSO." Neurocomputing 71.4-6 (2008): 1054-1060.

[4] Bansal, Jagdish Chand, and Kusum Deep. "A modified binary particle swarm optimization for knapsack problems." Applied Mathematics and Computation 218.22 (2012): 11042-11061.

[5] Ramos, Juan, et al. "An agent-based clustering approach for gene selection in gene expression microarray." Interdisciplinary Sciences: Computational Life Sciences 9.1 (2017): 1-13.

[6] Booma, P. M., and S. Prabhakaran. "Classification of genes for disease identification using data mining techniques." Journal of Theoretical and Applied Information Technology 83.3 (2016): 399.

[7] Natarajan, A., and R. Balasubramanian. "A Fuzzy Parallel Island Model Multi Objective Genetic Algorithm Gene Feature Selection For Microarray Classification." International Journal of Applied Engineering Research 11.4 (2016): 2761-2770.

[8] Bennet, Jaison,. "A hybrid approach for gene selection and classification using support vector machine." Int. Arab J. Inf. Technol. 12.6A (2015): 695-700.

[9] Nagpal, Rashmi, and Rashmi Shrivas. "Cancer Classification Using Elitism PSO Based Lezy IBK on Gene Expression Data." Journal of Scientific and Technical Advancements 1.4 (2015): 19-23.

[10] Thangaraju, Mr P., and R. Mehala. "Novel Classification based approaches over Cancer Diseases." system 4.3 (2015).

DOI 10.17148/IJARCCE.2021.10474

BIOGRAPHY

Name : Mr. M. Ananda Kumar **M.E., MBA.,**

Designation : Assistant Professor,
Department of Computer Science and Engineering,
Arasu Engineering College,
Kumbakonam

Specialization : Big Data, Cloud Computing, Cryptography, Data Mining, Distributed Computing, Machine Learning and Neural Networks.

Name : A. Manju **(M.E – CSE)**

College : Arasu Engineering College, Kumbakonam

Specialization : Big Data, Data Mining, Machine Learning.