TRANSCRIPTOME ANALYSIS OF PANCREATIC CANCER USING DEEP COMPUTATIONAL APPROACHES

A thesis submitted toward partial fulfillment of the requirements for the degree of

Master of Engineering in Biomedical Engineering

Submitted by

Purbanka Pahari
EXAM ROLL NO: **M4BMD18007**
CLASS ROLL NO: **001630201010**
REGISTRATION NUMBER: **137410** of **2016-2017**

Under the joint supervision of

Dr. Piyali Basak
Assistant Professor
School of Bioscience and Engineering
Jadavpur University

And

Dr. Anasua Sarkar
Assistant Professor
Department of Computer Science and Engineering
Jadavpur University

Course affiliated to

Faculty of Engineering and Technology
Jadavpur University
Kolkata-700032
India
2018

M.E. (Biomedical Engineering) course affiliated to
Faculty of Engineering and Technology
Jadavpur University
Kolkata-700032

CERTIFICATE OF RECOMMENDATION

This is to certify that the thesis entitled **"Transcriptome analysis of Pancreatic cancer using Deep Computational approaches"** is a bonafide work carried out by **PURBANKA PAHARI** under my supervision and guidance for partial fulfillment of the requirement for Post Graduate Degree of Master of Engineering in Biomedical Engineering during the academic session 2016-2018.

_____

THESIS ADVISOR
Dr. Anasua Sarkar
Assistant Professor
Department of Computer Science and Engineering
Jadavpur University
Kolkata-700032

_____

THESIS ADVISOR
Dr. Piyali Basak
Assistant Professor
School of Bioscience and Engineering
Jadavpur University
Kolkata-700032

_____

DIRECTOR
School of Bioscience and Engineering
Jadavpur University
Kolkata-700032

_____

Dean
Faculty Council of Interdisciplinary Studies, Law and Management
Jadavpur University
Kolkata – 700032

M.E. (Biomedical Engineering) course affiliated to
Faculty of Engineering and Technology
Jadavpur University
Kolkata-700032

Certificate of Approval**

The foregoing thesis is hereby approved as a creditable study of an Engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

_____ _____

Dr. Piyali Basak **Signature of Examiner**
(Thesis Advisor)
Assistant Professor
School of Bioscience and Engineering
Jadavpur University
Kolkata-700032

_____

Dr. Anasua Sarkar
(Thesis Advisor)
Assistant Professor
Department of Computer Science and Engineering
Jadavpur University
Kolkata-700032

** Only in case the thesis is approved.

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research work by the undersigned candidate, as part of her Master of Engineering in Biomedical Engineering studies during academic session 2016-2018.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by this rules and conduct, I have fully cited and referred all material and results that are not original to this work.

NAME: **PURBANKA PAHARI**

CLASS ROLL NO.: **001630201010**

EXAMINATION ROLL NO**.: M4BMD18007**

REGISTRATION NO.:**137410 of 2016-17**

THESIS TITLE: **Transcriptome analysis of Pancreatic cancer using Deep Computational approaches.**

SIGNATURE: **DATE:**

Acknowledgement

"Gratitude turns disappointment into lessons learned, discoveries made, alternatives explored, and new plans set in motion."

I express my deepest gratitude to my thesis advisor, Dr. Anasua Sarkar and Dr. Piyali Basak for giving me possible instructions and support, providing invaluable suggestions and creating practical opportunities to work in an interesting field of knowledge at every stage of this thesis work. The thesis would not have been completed without their constant supervision and help.

I would like to express my sincere gratitude and thanks to all respected teachers and PHD Scholars of School of Bio-Science and Engineering Jadavpur University for their constant help and encouragement for providing me all sorts of academic supports.

My sincere thanks go to all staff members of School of Bio-Science and Engineering, for their kind help and cooperation during the thesis work. I would like to thank my friends, especially Pratik Das, Nilotpal Das, Pamela Das, my senior Arijit Ghosh and all the well-wishers who were a constant source of my delight during the work span of the thesis. Last, but not the least, I must express my deep feelings for my parents and my sister who are the constant source of my energy, inspiration and determination for going ahead with my higher academic pursuit.

Date: ***Purbanka Pahari***

Dedicated to the Progress of Science and Research

Abstract

Pancreatic ductal adenocarcinoma (PDAC) is one of most forceful danger. The most widely recognized, pancreatic adenocarcinoma causes 85% of cases. The PDAC adenocarcinomas show up in the pancreas segment with stomach related chemicals. The more terrible visualization originates from the disappointment in discovery of PDAC prior because of poor remedial alternatives. The recognizable proof of Biomarker for PDAC is a continuous test. In this way, early indicative biomarkers and restorative focuses for PDAC are required to be recognized Earlier works demonstrate that epithelial to mesenchymal change [EMT] forms give helpful focuses to PDAC.KRAS, CDKN2A, TP53 and SMAD4 are the most noteworthy every now and again transformed qualities for PDAC. Crude CEL records of five microarray-based high-throughput data quality verbalization datasets (GSE28735, GSE15471, GSE41368, GSE32676 and GSE71989) containing enunciation data from through and through 105 normal pancreatic and 129 PDAC tissue tests were downloaded from the NCBI Quality Expression Omnibus (GEO). At that point the crude document information from each microarray are pre-handled in Matlab2017a. To choose those qualities which are significant and with minimum excess among them, we utilize progressive methodologies like Filter techniques and Normalization stage. We utilized essential part examination for highlight change and Fuzzy-C implies bunching and K-implies grouping for pre-handling. In this work, after pre-handling of the information, we have utilized three sorts of ghostly bunching techniques, Unnormalized, Ng-Jordan and proposed entropy-based Shi-Malik otherworldly grouping calculations to discover critical hereditary and organic data. There we have connected new Shannon's Entropy based separation measure to distinguish the groups on Pancreatic dataset. Next, we grouped utilizing Bagged Tree Ensemble technique and contrasted and different classifiers. The general accomplishment rate thus gained was normal of 96.48% for five testing datasets. Such a rate is 6– 15% higher than the looking at rates acquired by various existing DT (decision tree), DA (discriminant analysis) and SVM (support vector machines) and NN (nearest neighbor) approaches. At that point we utilized State full GRU Deep neural classifier and other Deep neural classifiers to analyze in Python. Here our accomplishment rate was normal 99.22% for five testing dataset which is higher than the looking at rates got by various existing state full and stateless LSTM and basic RNN approaches, deducing that the social event classifier is astoundingly reassuring and may transform into a huge test for biomarker recognizable proof. At long last, we think about the natural and useful connection of transcriptome profiles in our outcomes utilizing GO (Gene Ontology) Annotation instruments. Some Biomarkers are recognized through KEGG Pathway examination. The Biological examination and useful relationship of qualities in view of Gene Ontology terms demonstrate that the proposed strategy is useful for the determination of Biomarkers.

Contents

List of Figures

List of Tables

List of Abbreviation

PDAC	Pancreatic Ductal Adenocarcinoma
NCBI	National Center for Biotechnology Information
GEO	Gene Expression Omnibus
PCA	Principal Component Analysis
FCM	Fuzzy-C-Means Clustering
DNN	Deep Neural Network
ANN	Artificial Neural Network
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
DAVID	Database for Annotation Visualization and Integrated Discovery
BP	Biological Process
MF	Molecular Function
CC	Cellular Component
KNN	K-Nearest Neighbor
SVM	Support Vector Machine
LDA	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis
ROC	Receiver Operating Characteristic
CM	Confusion Matrix
PCP	Parallel Coordinates Plot
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network

CHAPTER 1

❖Introduction

1.1 Pancreatic Ductal Adenocarcinoma

1.2 Gene Expression

1.3 Materials

1.4 Principal Component Analysis

1.5 Clustering

1.6 Classification

1.7 Deep Neural Network

1.8 Biomarker

1.9 Functional Enrichment Analysis

1 Introduction

1.1 Pancreatic Ductal Adenocarcinoma

Pancreatic ductal adenocarcinoma is the most generally perceived sort of pancreatic threat, making up more than 80 percent of cases [1]. It is an epithelial tumor begins in the cells of the pancreas' channels, which transport juices containing basic stomach related proteins into the little digestive system [2]. In wellbeing, the pancreatic channel fills in as the course through which stomach related compounds and bicarbonate particle delivered in acinar cells achieve the small digestive system. Ductal cells and acinar cells together speak to the exocrine pancreas, from which most by far of pancreatic neoplasms emerge. It is presently trusted that the advancement of PDAC happens over a broadened timeframe, and likely takes after a stepwise movement like different carcinomas (colorectal carcinoma, specifically). This movement is described by the progress of an ordinary pancreatic conduit to a pre-intrusive antecedent injury known as pancreatic intraepithelial neoplasia (PanIN), which can at last form into an obtrusive PDAC [3]. This movement is impelled on by the progressive gathering of hereditary changes (Fig1).

The almost happening symptoms of PDAC are weight diminishment, anguish and fundamental clinical sign is jaundice. There are some more therapeutic Conditions, for instance, Chronic pancreatitis, serious pancreatitis, diabetes, cirrhosis, Helicobacter pylori sickness, human immunodeficiency disease (HIV) pollution, hepatitis B, cystic fibrosis, heftiness. Hereditary: Family history of pancreatic harm, Lynch issue, Li-Fraumeni issue various endocrine neoplasia 1, inborn chest and ovarian tumor, Familial atypical different mole melanoma issue, von-Hippel Lindau issue, Peutz-Jegher issue. Lifestyle: tobacco use, generous; alcohol use, high fat or cholesterol eat less which causes PDAC.

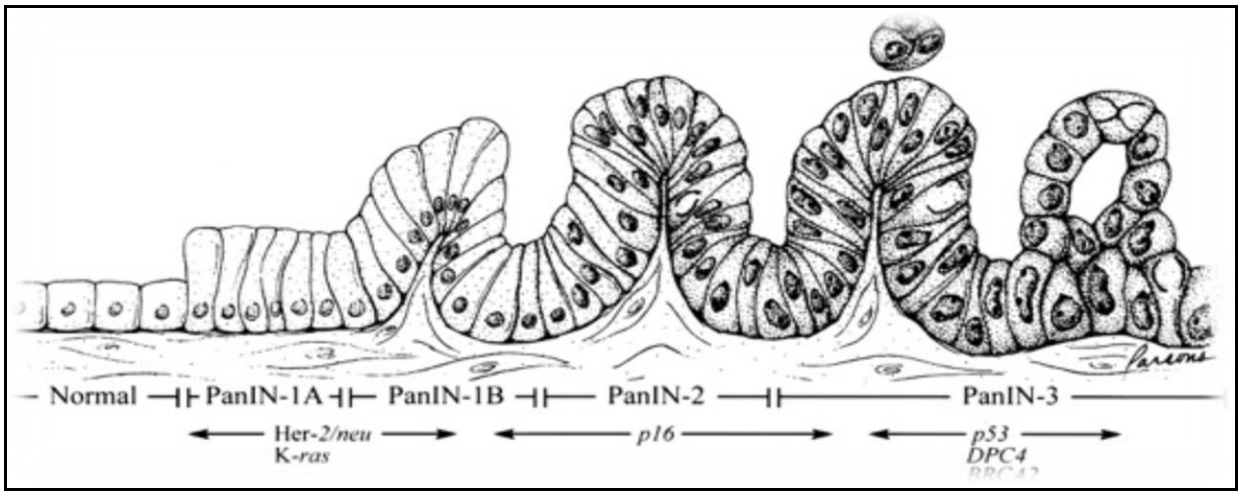


Fig1: The Progression Model for Pancreatic Cancer

Pancreatic ductal adenocarcinoma (PDAC) is an exceedingly intense risk. Till now, the patient's representation remains poor which, among others, is a direct result of the absence of strong 5early demonstrative biomarkers. Previously, contender explanatory biomarkers and therapeutic targets have been depicted from characteristics that were seen to be differentially conveyed in run of the mill versus tumor tests. As of late, new structures science approaches have been made to dismember quality enunciation data, which may yield new biomarkers. Keeping in mind the end goal to enhance our comprehension of the organic instruments basic PDAC, we investigated existing PDAC quality articulation datasets by applying an examination procedure to recognize enter qualities possibly engaged with the pathogenesis of PDAC.

1.2 Gene Expression

Gene expression is the procedure by which data from a gene is utilized as a part of the synthesis of a functional gene product. These products are frequently proteins, yet in non-protein coding qualities, for example, exchange RNA (tRNA) or little atomic RNA (snRNA) qualities, the item is a useful RNA.

A few stages in the quality articulation process might be tweaked, including the transcription, RNA splicing, translation, and post-translational change of a protein. Gene regulation gives the cell control over structure and work, and is the reason for cell separation, morphogenesis and the flexibility and versatility of any life form. Quality direction may likewise fill in as a substrate for transformative change, since control of the planning, area, and measure of quality articulation can profoundly affect the elements of the quality in a cell or in a multicellular organism.

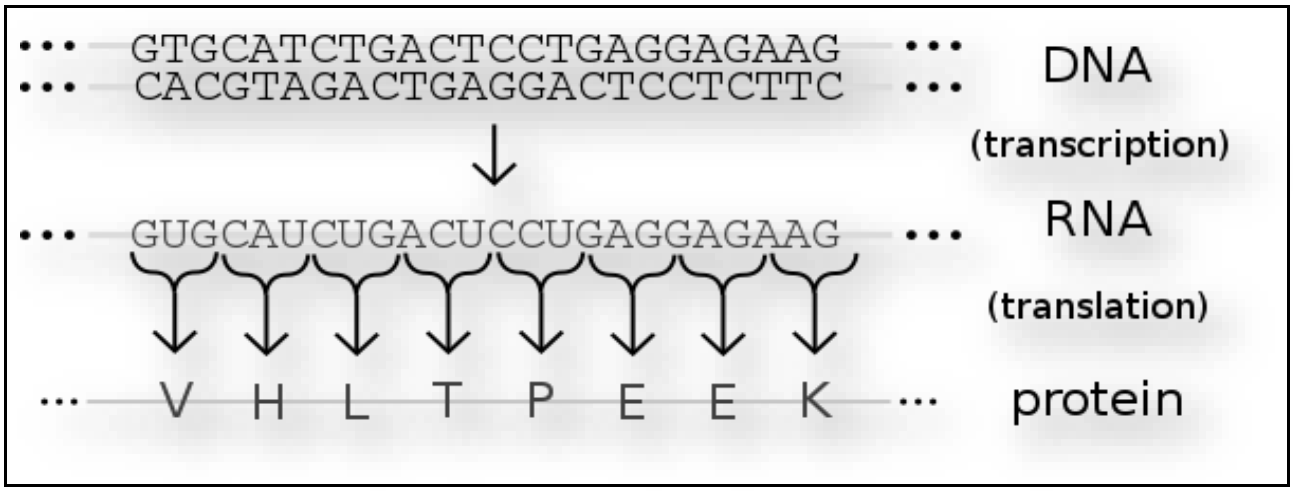


Fig2: Genes are expressed by being transcribed into RNA, and this transcript may then be translated into protein.

In genetics, gene articulation is the most key level at which the genotype offers ascend to the phenotype, i.e. detectable quality. The hereditary code put away in DNA is deciphered by quality articulation, and the properties of the articulation offer ascent to the living being's phenotype. Such phenotypes are frequently communicated by the combination of proteins that control the organism's shape, or that act as catalysts catalyzing specific metabolic pathways describing the organism. Direction of quality articulation is along these lines basic to an organism's improvement.

1.3 Materials

Crude CEL documents of five microarray-based transcriptome profile quality articulation datasets (GSE15471[4], GSE28735 [5], GSE32676 [6], GSE41368 [7], and GSE71989) containing articulation information from altogether 105 typical pancreatic and 129 PDAC tissue tests were downloaded from the NCBI Quality Expression Omnibus (GEO) (Table 1). High throughput data elaboration and figures presenting the results were obtained using the MATLAB 2017a and Python statistical environment.

Table 1: PDAC Gene Expression Dataset

Accession	Organism	Technology	Description	No of Normal Sample	No of PDAC Sample
GSE15471	Homo sapiens	Affymetrix Human Gene U133 Plus 2.0 Array	Whole-Tissue Gene Expression Study of Pancreatic Ductal Adenocarcinoma	36	36
GSE28735	Homo sapiens	Affymetrix Human Gene 1.0 ST Array	Microarray gene-expression profiles of pancreatic tumor and adjacent non-tumor tissues with pancreatic ductal adenocarcinoma	45	45
GSE32676	Homo sapiens	Affymetrix Human Gene U133 Plus 2.0 Array	Integrative Survival-Based Molecular Profiling of Human Pancreatic Cancer [mRNA]	7	25
GSE41368	Homo sapiens	Affymetrix Human Gene 1.0 ST Array	Combinatorial analysis of miRNA and mRNA expression in pancreatic ductal adenocarcinoma (PDAC)_mRNA	6	6
GSE71989	Homo sapiens	Affymetrix Human Gene U133 Plus 2.0 Array	The gene expression profiling of normal pancreatic and PDAC tissues.	8	14

1.4 Principal Component Analysis

PCA is a measurable strategy that uses an orthogonal transformation to change over an arrangement of perceptions of potentially related factors into an arrangement of estimations of straightly uncorrelated factors called principal components.

Principal component examination makes factors that are straight fusion of the first factors. The new factors have the property that the factors are overall orthogonal. The key parts can be utilized to discover bunches in an arrangement of information. PCA is a fluctuation centered approach looking to duplicate the aggregate variable change, in which parts reflect both normal and extraordinary difference of the variable. PCA is for the most part favored for motivations behind information lessening. Here we used principal component analysis for feature transformation.

PCA is numerically characterized as an orthogonal direct transformation that transforms the information to another organize framework with the end goal that the best difference by some projection of the information comes to lie on the primary arrange (called the first principal component), the second most prominent fluctuation on the second facilitate, and so on [8].

Consider a data matrix, X, with column-wise zero empirical mean (the sample mean of each column has been shifted to zero), where each of the n rows represents a different repetition of the experiment, and each of the p columns gives a particular kind of feature (say, the results from a particular sensor).

Mathematically, the transformation is defined by a set of p-dimensional vectors of weights or loadings $w_{(k)} = (w_1, \ldots, w_p)_{(k)}$ that map each row vector x_i of X to a new vector of principal component scores $t_{(i)} = (t_1, \ldots, t_m)_{(i)}$, given by

$$t_{k(i)} = x_{(i)} \cdot w_{(k)} \quad \text{for } i = 1, \ldots, n \text{ and } k = 1, \ldots, m$$

in such a way that the individual variables $t_1, \ldots, t_m$ of t considered over the data set successively inherit the maximum possible variance from X, with each loading vector w constrained to be a unit vector.

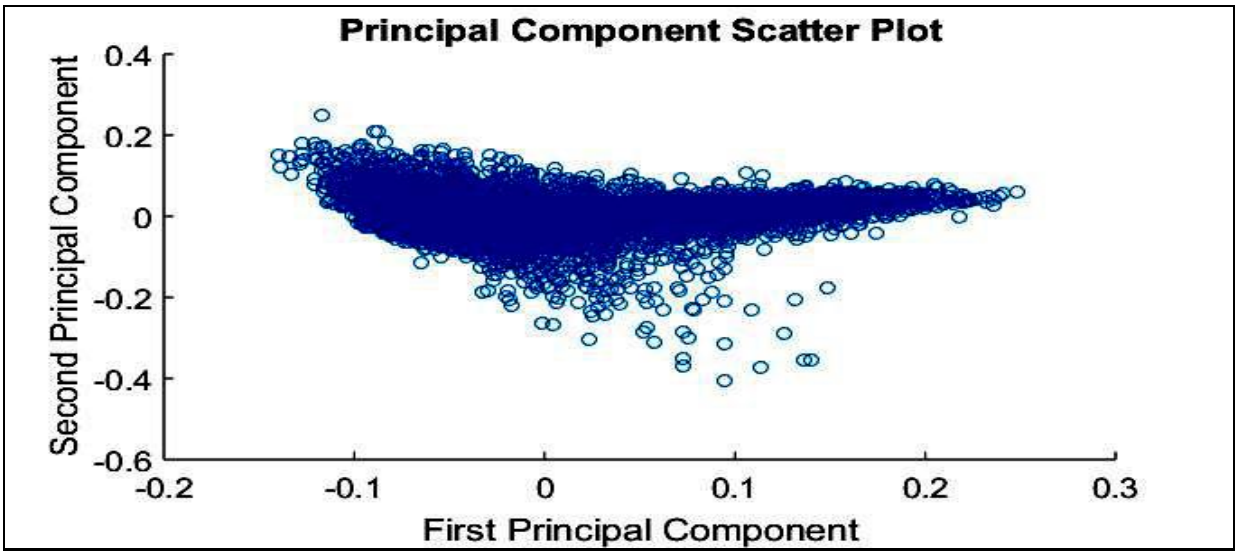


Fig3: Principal Component scatter plot of GSE28735 gene expression dataset

1.5 Clustering

Clustering is the way toward gathering information objects into an arrangement of disjoint classes, called clusters, so protests inside a class have high closeness to each other, while questions in independent classes are more divergent. Clustering is a case of unsupervised order. Grouping alludes to a technique that allots information articles to an arrangement of classes. Unsupervised implies that bunching does not depend on predefined classes and preparing illustrations while characterizing the information objects. Therefore, bunching is recognized from design acknowledgment or the territories of insights known as discriminant examination and choice investigation, which try to discover rules for grouping objects from a given arrangement of pre-ordered items.

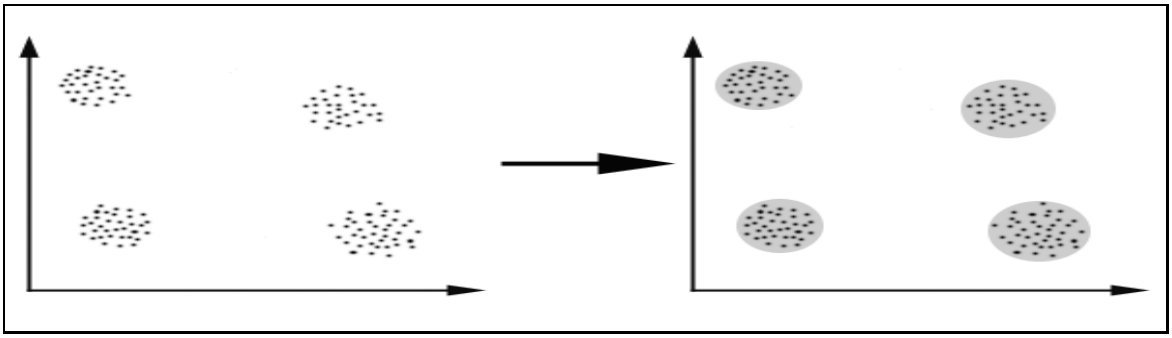


Fig4: Clustering of data points

In previous figure we easily identify the four clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are close according to a given distance

There are number of clustering methods are available to do the partitions. In our work, we have used three clustering algorithms are selected based on the problem definition.

- Fuzzy C-Means Clustering

FCM is a strategy for grouping which enables one bit of information to have a place with at least two clusters. This technique (created by Dunn in 1973 and enhanced by Bezdek in 1981) is as often as possible utilized as a part of example acknowledgment [9]. It depends on minimization of the accompanying target work:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 , 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij}^{m} is the degree of membership of x_i in the cluster j, x_i is the i^{th} of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij}^{m} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{2}{m-1}}}, c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

This iteration will stop when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of *Jm*.

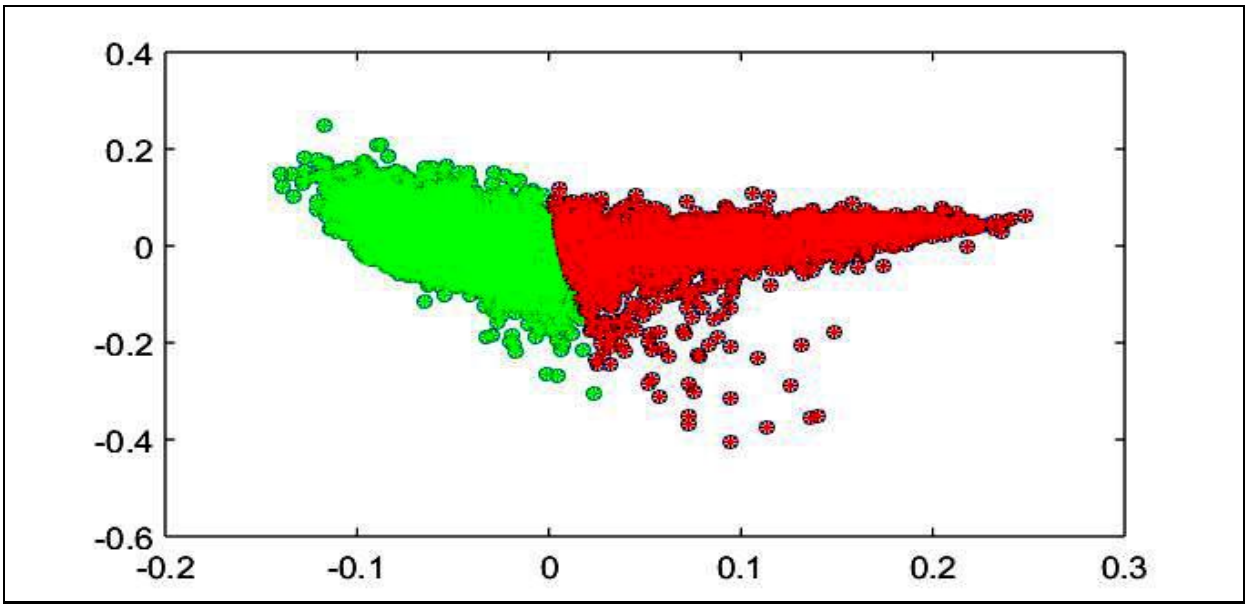


Fig5: Scatter plot of Fuzzy-c-means clustering solution of GSE28735 gene expression dataset

- K-Means Clustering

The K-means (MacQueen, 1967) is one of the most straightforward unsupervised learning calculations that tackle the outstanding bunching issue. The system takes after a straightforward and simple approach to order a given informational index through a specific number of groups (accept k clusters) settled from the earlier. The principle thought is to characterize k centroids, one for each group. These centroids ought to be put shrewdly in view of various area causes distinctive outcome. Thus, the better decision is to put them however much as could reasonably be expected far from each other. The subsequent stage is to take each guide having a place toward a given informational collection and partner it to the closest centroid. At the point when no point is pending, the initial step is finished, and an early groupage is finished. Now we have to re-compute k new centroids as barycenter of the bunches coming about because of the past advance. After we have these k new centroids, another coupling must be done between similar informational index focuses and the closest new centroid. A circle has been created. Because of this circle we may see that the k centroids change their area well-ordered until the point that no more changes are finished. At the end of the day, centroids don't move any more. At last, this calculation goes for limiting a goal work, for this situation a squared blunder work [10]. The objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 ,$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j ,is an indicator of the distance of the *n* data points from their respective cluster centers.

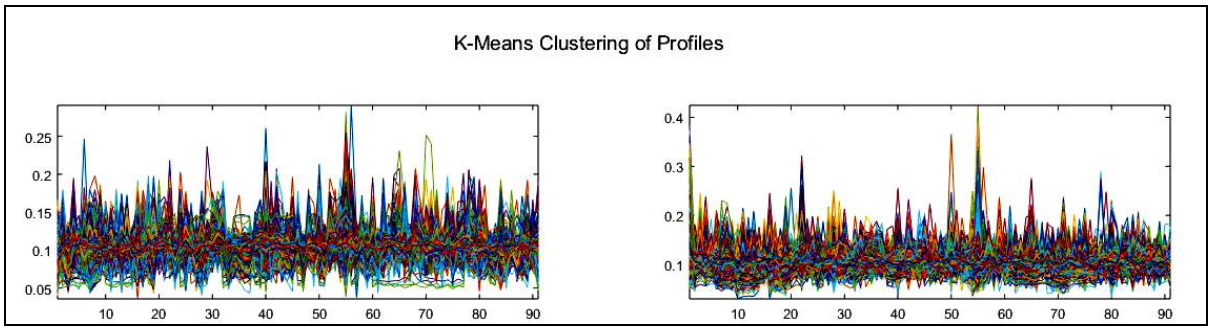


Fig6: K-Means Clustering of Profiles of GSE28735 gene expression dataset

- Spectral Clustering

In multivariate measurements and the grouping of information, otherworldly bunching systems make utilization of the range (eigenvalues) of the likeness lattice of the information to perform dimensionality diminishment before bunching in less measurements. The similitude lattice is given as an info and comprises of a quantitative evaluation of the relative comparability of each match of focuses in the dataset. We speak to the information as a full arrangement of pairwise similitudes. We usually get the similarity matrix using the Gaussian kernel method.

$$W(i,j) = \exp\left(-\left\| x(i) - x(j) \right\|^{\frac{2}{2\alpha^2}} \right)$$

This matrix can be represented as a graph where the weight of the edge represents the similarity between the two nodes [11].

- Graph Cut: Cut the graph in to n pieces while minimizing the number of edges (weight of edges) cut.
- Min-Cut: Find set of edges of minimal weight to disconnect the graph (tends to do lopsided cuts– e.g. cut off a vertex)
- Normalized Cut: $NormalizedCut(A,B) = \frac{cut(A,B)}{asso(A,V)} + \frac{cut(A,B)}{asso(B,V)}$

  where, $asso(A,V)$ = sum of all weights from nodes in A and

$cut(A, B)$ = sum of all weights between nodes in A and nodes in B

Given an enumerated set of data points, the similarity matrix may be defined as a symmetric matrix A, where Aij≥0 represents a measure of the similarity between data points with indices i and j. The general approach to spectral clustering is to use a standard clustering method on relevant eigenvectors of a Laplacian matrix of A.

There are a wide range of approaches to characterize a Laplacian which have distinctive numerical translations, thus the grouping will likewise have diverse elucidations. The eigenvectors that are significant are the ones that compare to littlest a few eigenvalues of the Laplacian except for the littlest eigenvalue which will have an estimation of 0. For computational effectiveness, these eigenvectors are regularly registered as the eigenvectors relating to the biggest a few eigenvalues of an element of the Laplacian.

1.6 Classification

In machine learning and measurements, classification is the issue of recognizing to which of an arrangement of classes another perception has a place, based on a preparation set of information containing perceptions whose classification enrollment is known. An arrangement procedure is a methodical way to deal with building grouping models from an info informational index [12]. Cases incorporate choice tree classifiers, nearest neighbor classifiers, ensemble classifiers, discriminant analysis classifiers, support vector machines, and neural network classifiers. The objective of arrangement is to precisely foresee the objective class for each case in the information. A calculation that executes characterization, particularly in a solid usage, is known as a classifier. The term classifier now and again likewise alludes to the scientific capacity, actualized by an order calculation, that maps input information to a classification.

1.7 Deep Neural Network

DNN is an artificial neural system (ANN) with different shrouded layers between the info and yield layers. It is general profound structure utilized for classification or regression. It enables nonlinear complex theory to be communicated.

An ANN depends on a gathering of associated units called counterfeit neurons, (comparable to axons in an organic cerebrum). Every association (neurotransmitter) between neurons can transmit a signal to another neuron. The getting (postsynaptic) neuron can process the signal(s) and after that flag downstream neurons associated with it. Neurons may have state, for the most part spoke to by genuine numbers, typically between 0 and 1. Neurons and neurotransmitters may likewise have a weight that shifts as learning continues, which can increment or abatement the quality of the flag that it sends downstream. Profound learning models are inexactly identified with data handling and correspondence designs in a natural sensory system, for example, neural coding that endeavors to characterize a connection between different jolts and related neuronal reactions in the brain [13].

Profound architectures incorporate numerous variations of a couple of fundamental methodologies. Every design has discovered accomplishment in specific domain. It isn't generally conceivable to think about the execution of various structures, unless they have been assessed on similar informational indexes.

DNNs are regularly feedforward organizes in which information streams from the information layer to the yield layer without circling back.

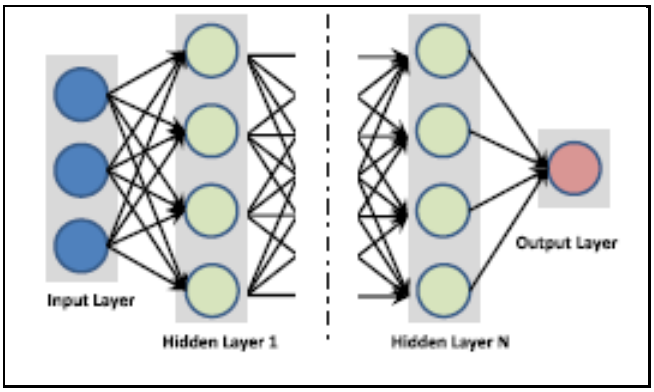


Fig7: Architecture of Deep Neural Network

1.8 Biomarker

The term biomarker, a portmanteau of natural marker, alludes to a general subcategory of restorative signs – that is, target signs of medicinal state observed from outside the patient – which can be estimated precisely and reproducibly. Restorative signs remain as opposed to medicinal manifestations, which are constrained to those signs of wellbeing or sickness saw by patients themselves.

National Institutes of Health Biomarkers Definitions Working Group characterized a biomarker as a trademark that is impartially estimated and assessed as a pointer of ordinary natural procedures, pathogenic procedures, or pharmacologic reactions to a helpful mediation.

A cancer biomarker is possibly one substance or a procedure which demonstrates that growth is available in the body. It can be characterized as a particle made from tumor. Identifying biomarkers in clinical space is yet a testing assignment [14]. Quality articulation exchanges quality data to union proteins, tRNA or snRNA.

Previously, it has e.g. been proposed that epithelial to mesenchymal progress [EMT] procedures may yield remedial focuses for PDAC and other strong tumors [15, 16]. KRAS, CDKN2A, TP53 and SMAD4 are the most every now and again changed qualities in PDAC, while a few different qualities are transformed at moderately low frequencies [17]. Likewise, a few articulation related biomarkers at both the RNA and protein level with symptomatic, prognostic and prescient esteem have been recognized [18], including starch antigen 19– 9 (CA19– 9) and carcinoembryonic antigen (CEA, as of now named CEACAM5) [17], while different biomarkers have been prohibited as PDAC prognosticators [19].

1.9 Functional Enrichment Analysis

- Gene Ontology

Gene Ontology (GO) is a noteworthy bioinformatics activity to bind together the portrayal of gene and gene product properties over all species [20]. GO is the system for the model of science. The GO venture was set up to give a typical dialect to portray parts of a gene product's science. The utilization of a reliable vocabulary enables qualities from various species to be compared based on

their GO annotations. The goal of GO is to give controlled vocabularies to the depiction of the biological process, molecular function, and cellular component of gene products. These terms are to be utilized as traits of quality items by organism databases, encouraging uniform questions crosswise over them. The controlled vocabularies of terms are organized to enable annotation of gene products to GO terms at different levels of detail and to inquiry for quality items that are associated with comparative procedures, capacity and parts.

Ontologies

- Molecular Function

Molecular functions terms depict exercises that happen at the sub-atomic level, for example, reactant movement or restricting action. GO Molecular functions terms speak to exercises as opposed to the substances (particles or edifices) that play out the activities, and don't determine where, when, or in what setting the move makes put. Molecular functions for the most part compare to exercises that can be performed by individual gene products, however a few exercises are performed by collected buildings of gene products. Cases of wide practical terms are synergist movement and transporter action; cases of smaller utilitarian terms are adenylated cyclase action or Toll receptor official.

It is anything but difficult to befuddle a quality item name with its sub-atomic capacity; thus, GO sub-atomic capacities are regularly annexed with the word action.

- Cellular Component

These terms depict an area, in respect to cell compartments and structures, involved by a macromolecular machine when it completes a sub-atomic capacity. There are two manners by which scholars portray areas of quality items: (a) with respect to cell structures (e.g., cytoplasmic side of plasma layer) or compartments (e.g., mitochondrion), and (b) the stable macromolecular edifices of which they are parts (e.g., the ribosome). Dissimilar to alternate parts of GO, cell segment ideas allude not to forms yet rather a cell life structures.

- Biological Process

A biological procedure term depicts a progression of occasions achieved by at least one sorted out gatherings of molecular functions. Cases of wide natural process terms are cell physiological process or signal transduction. Cases of more particular terms are pyrimidine metabolic process or alpha-glucoside transport. The general run to help with recognizing an organic procedure and an atomic capacity is that a procedure must have more than one unmistakable advances.

- Pathway Analysis

GO annotations is the model of science. Annotations are articulations portraying the elements of genes, utilizing ideas in the Gene Ontology. The least difficult and most basic explanation joins one quality to one capacity, e.g. FZD4 + Wnt flagging pathway.

In bioinformatics explore, pathway investigation programming is utilized to distinguish related proteins inside a pathway or building pathway all over again from the proteins of premium. This is useful when considering differential articulation of a quality in an illness or breaking down any omics dataset with countless. By looking at the adjustments in quality articulation in a pathway, its organic causes can be investigated. Pathway is the term from atomic science which delineates a fake rearranged model of a procedure inside a cell or tissue. Pathway examination comprehends or decipher omics information from the perspective of sanctioned earlier learning organized as pathways graphs. It permits finding unmistakable cell forms (Cellular procedures), sicknesses or flagging pathways that are measurably connected with determination of differentially communicated qualities between two samples [21]. Frequently however mistakenly pathway examination is utilized as equivalent word for arrange investigation (utilitarian enhancement examination and quality set analysis) [22].

Like KEGG (Kyoto Encyclopedia of Genes and Genomes) is an accumulation of databases managing genomes, organic pathways, ailments, medications, and synthetic substances. KEGG is used for bioinformatics research and training, incorporating information investigation in genomics, metagenomics, metabolomics and different omics studies, demonstrating and reenactment in frameworks science, and translational research in tranquilize advancement.

CHAPTER 2

❖Literature Review

2.1 Weighted gene co-expression network on pancreatic ductal adenocarcinoma development

2.2 Cluster Analysis for Gene Expression Data

2.3 Evaluation of Gene Expression Classification Studies

2.4 An advanced cancer type classifier based on deep learning

2. Literature Review

2.1 Weighted gene co-expression network analysis reveals key genes involved in pancreatic ductal adenocarcinoma development

Reason Pancreatic ductal adenocarcinoma (PDAC) is a very forceful threat. Up till now, the patient's forecast stays poor which, among others, is because of the lack of solid early indicative biomarkers. Previously, applicant symptomatic biomarkers and remedial targets have been depicted from qualities that were observed to be differentially communicated in typical versus tumor tests. As of late, new frameworks science approaches have been produced to dissect quality articulation information, which may yield new biomarkers [23].

Strategies PDAC microarray-based quality articulation datasets, recorded in the Gene Expression Omnibus (GEO) database, were broke down. After pre-preparing of the information, Matteo Giulietti, Giulia Occhipinti et all fabricated two last datasets, Normal and PDAC, enveloping 104 and 129 patient examples, separately. Next, they developed a weighted quality co-articulation arrange and recognized modules of coexpressed qualities recognizing ordinary from sickness conditions. Practical explanations of the qualities in these modules were completed to feature PDAC-related sub-atomic pathways and regular administrative systems. At last, general survival examinations were completed to evaluate the reasonableness of the qualities recognized as prognostic biomarkers. Results Using WGCNA, they distinguished a few key qualities that may assume critical parts in PDAC. These qualities are essentially identified with either endoplasmic reticulum, mitochondrion or film capacities, display transferase or hydrolase exercises and are associated with organic procedures, for example, lipid digestion or transmembrane transport. As an approval of the connected technique, they found that a portion of the distinguished key qualities (CEACAM1, MCU, VDAC1, CYCS, C15ORF52, TMEM51, LARP1 and ERLIN2) have beforehand been accounted for by others as potential PDAC biomarkers. Utilizing general survival investigations, they found that few of the recently recognized qualities may fill in as biomarkers to stratify PDAC patients into low-and high-hazard gatherings. They found that two modules of coexpressed qualities varied fundamentally between the Normal and PDAC systems, proposing a part in the pathogenesis of PDAC. Hence, they limited the rundown of qualities inside these modules by distinguishing just the center qualities, i.e., the most PDAC-related qualities as

indicated by WGCNA. Useful advancement investigation of these qualities uncovered that they are identified with either endoplasmic reticulum (ER), mitochondrion or film capacities, show transferase or hydrolase exercises, and are identified with natural procedures, for example, lipid digestion or transmembrane transport. Conclusions Using this new framework science approach, they distinguished a few qualities that seem, by all accounts, to be basic to PDAC improvement. Thusly, they may speak to potential analytic biomarkers and helpful focuses with clinical utility [24].

2.2 Cluster Analysis for Gene Expression Data

DNA microarray innovation has now made it conceivable to at the same time screen the articulation levels of thousands of qualities amid vital organic procedures and crosswise over accumulations of related examples. Explaining the examples covered up in quality articulation information offers a gigantic open door for an improved comprehension of practical genomics. However, the expansive number of qualities and the multifaceted nature of organic systems enormously builds the difficulties of grasping and deciphering the subsequent mass of information, which frequently comprises of a great many estimations. An initial move toward tending to this test is the utilization of bunching systems, which is fundamental in the information mining procedure to uncover common structures and distinguish intriguing examples in the basic information. Numerous customary clustering algorithms have been adjusted or straightforwardly connected to quality articulation information, and new algorithms have as of late been proposed particularly going for quality articulation information. These clustering algorithms have been demonstrated valuable for distinguishing naturally significant gatherings of qualities and tests [25].

2.3 Evaluation of Gene Expression Classification Studies

Classification strategies utilized as a part of microarray examines for quality articulation are differing in the way Putri W. Novianti, Kit C. B. Roes, Marinus J. C. Eijkemans et all manage the hidden many-sided quality of the information, and additionally in the system used to manufacture the order demonstrate. The MAQC II consider on malignancy arrangement issues has discovered that execution was influenced by variables, for example, the characterization calculation, cross approval technique, number of qualities, and quality choice strategy. In this paper, they consider

the speculation that the illness under examination fundamentally figures out which strategy is ideal, and that also test measure, class awkwardness, kind of medicinal inquiry (symptomatic, prognostic or treatment reaction), and microarray stage are possibly powerful. An orderly writing audit was utilized to separate the data from 48 distributed articles on non-disease microarray grouping contemplates. The effect of the different factors on the detailed arrangement precision was broke down through arbitrary catch calculated relapse. The sort of medicinal inquiry and strategy for cross approval overwhelmed the clarified variety in exactness among examines, trailed by infection class and microarray stage. Altogether, 42% of the between think about variety was clarified by all the examination and issue particular factors that they contemplated together.

The measurable investigation of microarray information might challenge, with the inalienable danger of finding a false positive outcome because of the high dimensional nature of the information. Basic defects in the three distinctive objectives for the factual examination of microarray information (e.g. differential articulation, class revelation (unsupervised), and class expectation (supervised)) have been discovered [26]. Irregularity in the aftereffects of microarray examinations inside the same dataset tragically has likewise been accounted for, particularly for class forecast [27]. The inconstancy of the announced order accuracies might be because of the variety in the strategies used to construct the characterization show, e.g. the sort of characterization display, cross approval and quality choice system [28]. Furthermore, the execution of a prescient model may likewise rely upon normal for the microarray information [29].

They searched microarray quality articulation considers through PubMed (US National Library of Medicine National Institute of Health) for pertinent papers. Connected examinations in which the specialist planned to assemble directed models considering microarray quality articulation trial information were basically of intrigue.

Some classification strategies can naturally deal with the scourge of dimensionality (p>>n), yet others require a quality determination advance to achieve a lower measurement before applying the order technique. A portion of the examinations chose qualities univariately considering a measurement passing an edge for determination or the best K qualities to sustain the classifier. In different examinations, the quality choice strategy was gone for finding an ideal arrangement of qualities by stepwise emphasizing amongst choice and classifier building. In this manner, they gathered the quality determination strategy considering their connection with the classifier, specifically channel (e.g. univariate determination), wrapper (e.g. stepwise advancement of the

chose quality set) and inserted (e.g. punished probability relapse). Gathering was likewise done on the characterization technique into two classifications, contingent upon their capacity to identify collaborations between qualities. Qualities can be initiated freely yet in addition be actuated through the enactment of different qualities. Because of this marvel, the grouping techniques that can naturally display collaborations are relied upon to have preferable execution over the individuals who can't, in any event in a few investigations. The techniques that could distinguish the collaboration (alluded to as "association classifiers") in our audit were tree-based strategies, calculated relapse, support vector machines (SVM), k-Nearest Neighbors (KNN), artificial neural network (ANN), and weighting voting strategies. In the interim, discriminant investigation, prediction analysis of microarray (PAM), compound covariate indicator, closest centroid was ordered into the gathering of techniques that couldn't naturally recognize connections. The exactness of arrangement models considering gene expression microarray information relies upon think about and issue in specific elements. The cross-approval procedure has a vital effect in clarifying the inconstancy over the investigations.

2.4 An advanced cancer type classifier based on deep learning

With the improvements of DNA sequencing innovation, a lot of sequencing information have wind up accessible as of late and give remarkable chances to cutting edge affiliation thinks about between physical point changes and malignancy subtypes. However, in existing strategies, issues like high information sparsity, little volume of test estimate, and the use of straightforward direct classifiers, are significant hindrances in enhancing the grouping execution. To address the deterrents in existing studies, Yuchen Yuan, Jinman Kim, David Dagan Feng et all propose Deep Gene, a propelled deep neural network (DNN) based classifier, that comprises of three stages: right off the bat, clustered gene filtering (CGF) concentrates the quality information by change event recurrence, sifting through the lion's share of unessential qualities; besides, the index sparsity reduction (ISR) changes over the quality information into files of its non-zero components, subsequently fundamentally smothering the effect of information sparsity; at long last, the information after CGF and ISR is bolstered into a DNN classifier, which removes abnormal state highlights for precise grouping. Trial comes about on our curated TCGA-Deep Gene dataset, which is a reformulated subset of the TCGA dataset containing 12 chose sorts of malignancy, demonstrate that CGF, ISR and DNN all contribute in enhancing the general arrangement

execution. The DNN classifier is the backbone of Deep Gene, which conducts the classification and produces the last yield. In this paper, they observe the critical preferred standpoint of Deep Gene against three generally embraced classifiers, among which

Deep Gene shows no less than 24% of execution change. To analyze the execution of the DNN classifier itself without the pre-preparing ventures of CGF and ISR, they likewise record the precision of the DNN classifier with crude quality information, which has demonstrated that the DNN classifier still produces the best precision (60.1% against the second best 52.7% of SVM). To additionally approve that the 10-overlap approval precision of DNN is without a doubt higher than that of SVM, they expect that these two classifiers are autonomous of each other, what's more, channel t-test with the invalid theory that these two classifiers have break even with approval exactness under the significance of 0.001.

They lead probes the aggregated TCGA Deep Gene dataset, which is a reformulated subset of the TCGA dataset with changes on 12 kinds of malignancy. Controlled variable analyses show that CGF, ISR and DNN classifier all have critical commitment in enhancing the arrangement exactness. They compare Deep Gene with three broadly embraced information classifiers, the aftereffects of which display the momentous points of interest of Deep Gene, which has accomplished > 24% of execution change as far as testing precision against the comparison classification techniques. They showed the favorable circumstances and possibilities of the Deep Gene demonstrate data processing, and they recommend that the model can be stretched out and exchanged to another complex genotype phenotype.

Based on profound learning and physical point change information, they devise Deep Gene, a propelled malignancy type classifier, which tends to the impediments in existing studies. Examinations show that Deep Gene beats three broadly embraced existing classifiers, which is for the most part ascribed to its profound learning module that can separate the high-level highlights between combinatorial physical point changes and cancer types [30].

CHAPTER 3

❖Entropy based Spectral Clustering of Gene Expression

3.1 Preface

3.2 Methodology

3.3 Experimental Framework

3.4 Result & Discussion

3. Entropy based Spectral Clustering of Gene Expression

3.1 Preface

Pancreatic ductal adenocarcinoma (PDAC) is one of most forceful harm. The distinguishing proof of Biomarker for PDAC is a continuous test. The high dimensional PDAC gene expression dataset in Gene Expression Omnibus(GEO) database, is analyzed in this work. To choose those qualities which are important and in addition with minimum excess among them, we utilize progressive methodologies like Filter techniques and Normalization stage. In this work, after pre-preparing of the information, we have utilized three sorts of spectral clustering strategies, Unnormalized, Ng-Jordan and proposed entropy-based Shi-Malik spectral clustering calculations to discover imperative hereditary and organic data. There we have connected new Shannon's Entropy based separation measure to distinguish the clusters on Pancreatic dataset. Some Biomarkers are recognized through KEGG Pathway examination. The Biological investigation and practical relationship of qualities in view of Gene Ontology(GO) terms appear that the proposed technique is useful for the choice of Biomarkers.

3.2 Methodology

105 normal pancreatic and 129 PDAC tissue samples are obtained from GSE28735 gene expression dataset of NCBI Gene Expression Omnibus (GEO).

- Spectral Clustering

Spectral clustering exploits eigenvalues of similarity matrix of data for dimensionality reduction. Three distinctive spectral clustering approaches, which we have utilized as a part of our analyses, have been clarified beneath:

- Unnormalized spectral clustering

Input: Similarity matrix $S \in R^{n \times n}$, k clusters

i. To generate similarity graph. Let W = weighted adjacency matrix.

ii. Calculate unnormalized Laplacian L.

iii. Calculate first k eigenvectors in $V \in R^{n \times k} = \{v_1 ... v_k\}$ of L.

iv. For $i = 1...\eta th$ row, let $V_i \in R^k$

v. Cluster points $(V_i), i = 1...n$ with k-means method to $C_1 ... C_k$ clusters.

Output: *k* Clusters, $C_1 ... C_k$ [31].

- Spectral Clustering method by Shi and Malik (2000)

Input: Similarity matrix $S \in R^{n \times n}$, *k* clusters

i. Generate similarity graph. Let W = weighted adjacency matrix.

ii. Calculate unnormalized Laplacian L.

iii. Calculate first *k* eigenvectors in $V \in R^{n \times k}$ of generalized Eigen equation $Lv = \lambda Dv$.

iv. For $i = 1...\eta th$ row, let $V_i \in R^k$.

v. Cluster points $(V_i), i = 1...n$ with k-means method to $C_1 ... C_k$ clusters.

Output: *k* Clusters, $C_1 ... C_k$ [32] [33].

- Normalized method of spectral clustering by Ng, Jordan, and Weiss (2002)

Input: Similarity matrix $S \in R^{n \times n}$, *k* clusters

i. Generate similarity graph. Let W = weighted adjacency matrix.

ii. Calculate normalized Laplacian L_{norm}.

iii. Calculate first k eigenvectors in $V \in R^{n \times k}$ of L_{norm}.

iv. Create row normalized matrix $U \in R^{n \times k}$ from V as $u_{ij} = v_{ij} \left(\sum_k v_{ik}^2 \right)^{1/2}$.

v. For $i = 1...\eta th$ row, let $V_i \in R^k$.

vi. Cluster points $(V_i), i = 1...n$ with k-means method to $C_1 ... C_k$ clusters.

Output: *k* Clusters, $C_1 ... C_k$ [34].

These three methods are similar, yet they exploit three separate graph Laplacians. These algorithms change the abstract data points x_i to points $V_i \in R^k$. These changes of graph Laplacians enhance the clusters detection among data differently.

- Shannon's Entropy theory

Shannon in 1948[35] defined the entropy of probability density function p(x) as

$$H = -\int p(x) \ln p(x) dx$$

3.3 Experimental Framework

Our proposed framework comprises of two stages - beginning personality framework and Eigen vectors generation stage and after that new Shi Malik Spectral clustering with Shannon's Entropy is utilized. At first, we input our picked Pancreatic tumor quality articulation dataset for over-articulation and little fluctuation filtering and normalization technique in MATLAB 2015a and at that point apply Shi-Malik Spectral clustering algorithm as portrayed previously. In this stage, in the wake of creating the information we build the identity matrix W and diagonal matrix D. After that, we frame U Matrix utilizing k Eigenvectors. In next stage, we register new Shannon's Entropy proposed distance measure to process cluster allotments for genes. In K-implies clustering advance, for getting K group centroid areas in the K-by-P distance framework for P no of genes, we characterize new distance measure utilizing Shannon's Entropy as characterized in the following condition:

$$d_{SE}\left(x_i, c_k\right) = \frac{\left[d_e\left(x_i, c_k\right) \times \left|H\left(x_i\right) - H\left(c_k\right)\right|\right]}{\sum_{\forall x_j \in c_{k \neq i \neq j}} d_e\left(x_i, c_k\right)}$$

where x_i denotes i^{th} point and c_k , k 1...m denotes relevant cluster. The quantitative evaluation with validity indices are performed on the entropy-based solution to show efficiency of the new method.

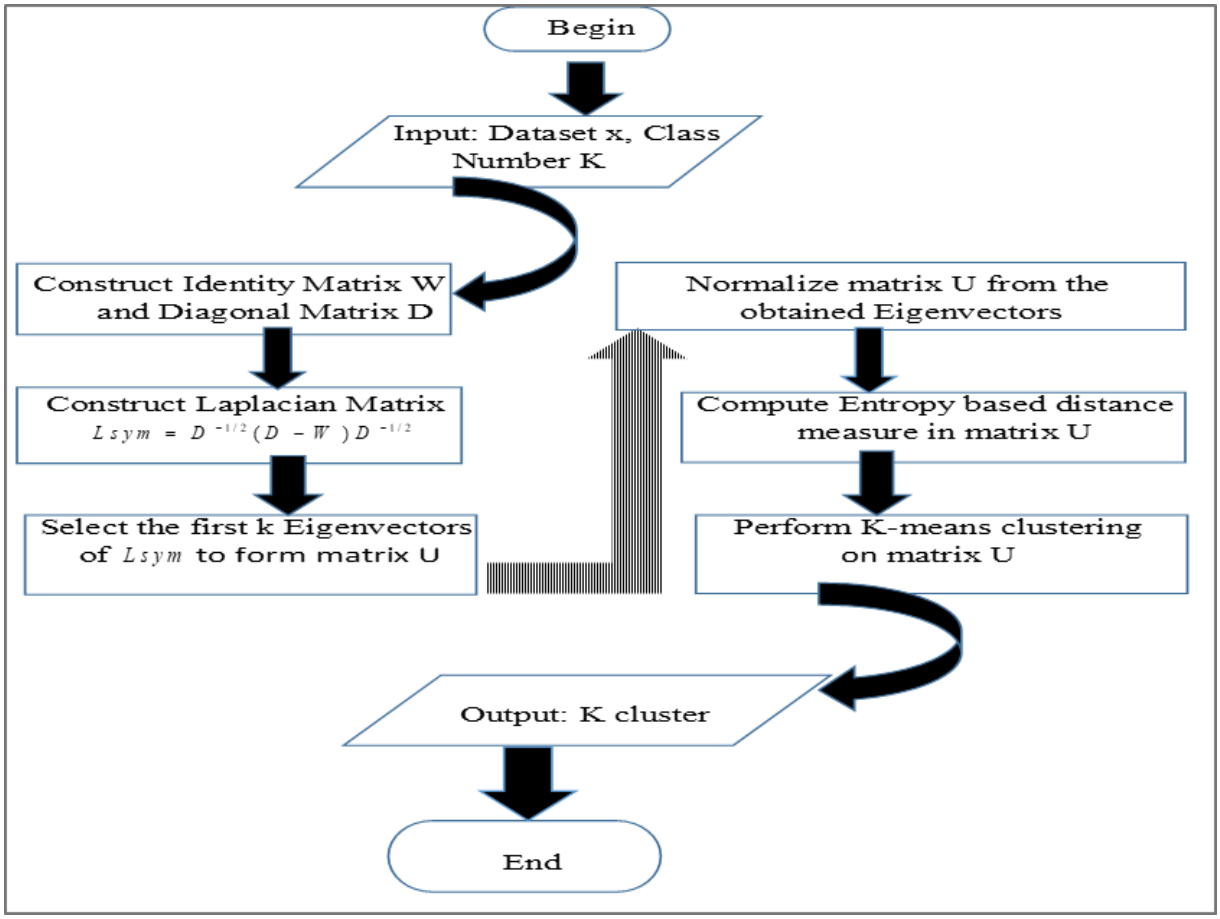


Fig8: The Framework of Proposed Algorithm

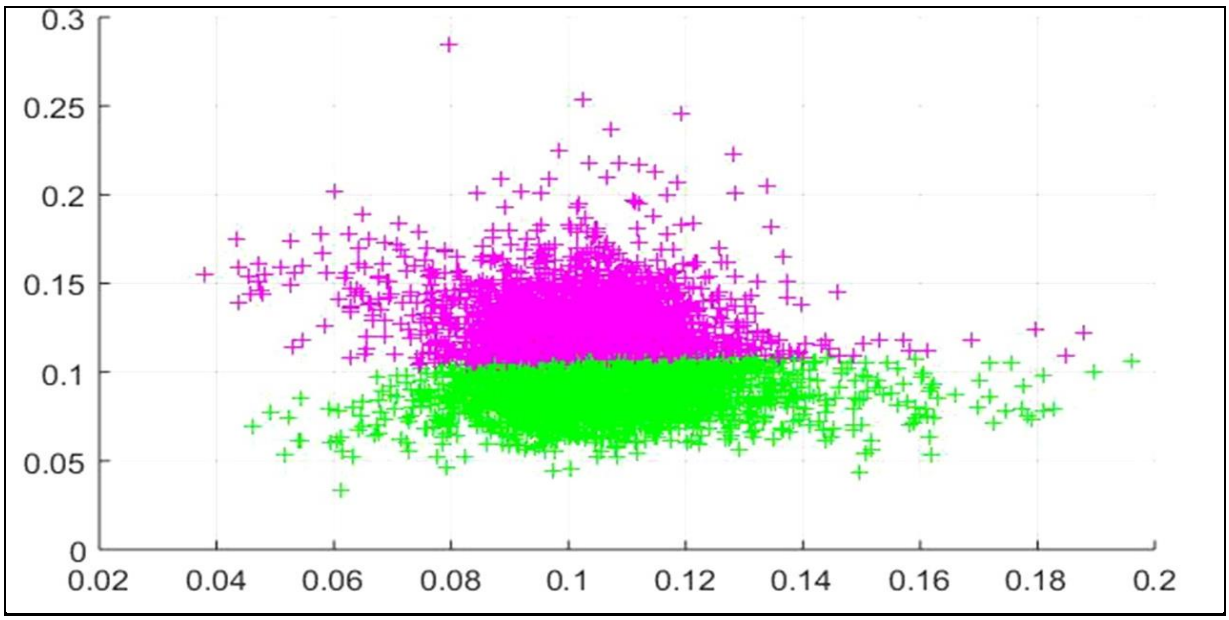


Fig9: Cluster profiles using Entropy distance proposed algorithm

3.4 Result & Discussion

We register two known Unnormalized and Ng-Jordan-Weiss Spectral clustering algorithms first and we at that point apply our proposed algorithm – modified Shi-Malik Spectral clustering utilizing Entropy based distance as depicted in past area to get the cluster arrangements over picked Pancreatic malignancy quality articulation dataset. After getting the cluster arrangements, we approve the outcomes utilizing Davies-Bouldin and Dunn validity indices. Moreover, we dissect the outcomes statistically and finally distinguished the Biomarkers utilizing GO annotations and KEGG pathways in DAVID (Database for Annotation Visualization and Integrated Discovery) tools. Figure 2 demonstrates the two cluster profiles which we get as cluster arrangement after applying our proposed algorithm.

- Validity Analysis

Table 2: Validity indices values for different algorithms

Index	**Unnormalized Spectral Clustering**	**Normalized Spectral Clustering**	**Proposed Shi Malik Spectral Clustering with Shannon's Entropy**
Davies Bouldin Index	1.8938	2.4505	1.1422
Dunn Index	0.4975	0.8042	1.7227

The acquired clustering arrangements are then assessed quantitatively by limiting Davies-Bouldin and expanding Dunn indices. From Table 2, we get least Davies Bouldin esteem as 1.1422 and most extreme Dunn esteem as 1.7227 in Shi Malik with our new proposed algorithm contrasted with other two known algorithms.

- Statistical Analysis

A noteworthy non-parametric statistical investigation is known as Wilcoxon's rank sum for autonomous examples with 5% importance level. Davies Bouldin record esteems and Dunn list esteems over and again deliver by 10 back to back keeps running of Unnormalized Spectral Clustering, Shi Malik Spectral clustering with Shannon's entropy algorithm, Normalized Ng. Jordan Spectral clustering. We got the outcome from the median of each gathering. It is observed that Shi Malik Spectral clustering with Shannon's entropy algorithm provides better median values than others as appeared in Table 3.

Table 3: Median values of Davies Bouldin and Dunn indices after performing ten times on different algorithms

Data	**Validity Index**	**Algorithm**		
		Unnormalized Spectral Clustering	**Normalized Spectral Clustering**	**Proposed Shi Malik Spectral Clustering with Shannon's Entropy**
PDAC	DB	1.9976	2.4505	1.1422
	Dunn	0.5078	0.8042	1.7227

Table 4: P values of applied two known algorithms comparing with our proposed algorithm by using Wilcoxon's rank sum test

Algorithm	**Comparison with modified Shi Malik Spectral clustering with Shannon's entropy-based distance algorithm**	
	H value	**P value**
Unnormalized Spectral Clustering	1	4.01E-13
Normalized Ng Jordan Spectral Clustering	1	1.13E-11

In Table 4 we demonstrate the H esteems and P estimations of spectral clustering and Ng. Jordan spectral clustering contrasting and our proposed Shi-Malik spectral clustering with Shannon's entropy algorithm by utilizing Wilcoxon's rank sum test. P values in the two cases are little which is demonstrating 5% importance level.

- Gene Ontology Analysis

We utilized DAVID device for acquiring GO comment. Here we can watch the organic elements of some biomarker qualities. We demonstrate some acquired GO-TERM with p esteem $p<0.05$ and with high % of quality in Table 5.

Table 5: GOTERM for genes from our cluster solutions of proposed method

Group	**GO Term**		**Ontology Description**	**Percentage of presence**	**Genes**
Cluster1 (Cancer)	BP	GO:0016043	cellular component organization	18.08163	TGFB3, CDH13, MARK1, JUP, RCC2
		GO:0048513	organ development	10.65306	TGFBR3, PLAU, SMAD3, TP53
	MF	GO:0005488	binding	70.77551	MAP3K6, JUP, GNA13, TGFB2
		GO:0000166	Nucleotide binding	14.86	MAP3K6, KRAS, TP53, ADCY4

	CC	GO:0044446	intracellular organelle part	25.918	MAPK3, TP53, MAPK8, RCC2
		GO:0005737	cytoplasm	45.87	MAPK3, MAPK8, LDAH, JUP
Cluster2 (Normal)	BP	GO:0044765	Single organism transport	17.95186	RPL17, SCN3B, VTCN1, LTBP4
		GO:1902578	single organism localization	18. 64545	RPL17, SCN3B, VTCN1, LTBP4
	MF	GO:0043167	ion binding	20.60384	RORC, ZNF253, CDH23, FAH
		GO:0046872	metal ion binding	19.33905	RORC, ZNF253, CDH23, ALPPL2
	CC	GO:0044421	extracellular region part	20.39984	LDHB, LTBP4, NAA16, GNG7

		GO:0031982	vesicle	19.13505	LDHB, LTBP4, NAA16, GNG7

We study about the natural and useful connection of qualities in our outcomes utilizing GO (Gene Ontology) Annotation instruments. In Table 5 we express some GO Terms for Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) sub ontologies with p esteem <0.05. The level of essence infers the qualities exhibit in GO from all genes. The observation of GO Terms demonstrate that the clustering result is improved with different procedures including cell segment official, nucleotide binding, organ improvement. Prior in [36], comparative GO MF terms, to be specific GO:0005488 and GO:0000166 are moreover disclosed to relate to Pancreatic tumor. We discovered a few qualities from GO explanations which are generally in charge of tumor, for example, LAMB3, MAP3K6, MMP13, LPHN2, MAP4K5, CD3G, B2M, ITGB3, MMP2 and so forth. Among them a few qualities we have effectively found in others work [36].

Table 6: KEGG pathway analysis on clustering solutions of proposed method

Group	**KEGG Pathway**	**Pathway Name**	**P value**	**Genes**
Cluster1 (Cancer)	hsa05212	Pancreatic cancer	0.022877	RASS5, GNA13, ADCY7, SMAD3, RAD51, AKT3, CHUK, PIK3R3
	hsa05200	Pathways in cancer	0.002121	SLC2A1, AKT3, CHUK, RAD51,

				STAT5A, NFKB2, ADCY4
	hsa05230	Central carbon metabolism in cancer	0.040573	PIK3R3, KRAS, MAPK3, AKT3, SLC2A1
	hsa05223	Non-small cell lung cancer	0.028272	GRB2, AKT3, KRAS, PIK3R3
	hsa05221	Acute myeloid leukemia	0.0127	CHUK, AKT3, MAPK3, GRB2, KRAS, PIK3R3
Cluster2 (Normal)	hsa04975	Fat digestion and absorption	1.19E-04	PLA2G5, FABP1, DGAT2, APOB
	hsa03010	Ribosome	0.002085	MRPS15, RPL13, RPL36
	hsa00640	Propionate metabolism	3.87E-04	ALDH6A1, MCEE,

				LDHB, ABAT
	hsa04911	Insulin secretion	0.003137	TRPM4, RYR2, INS, ABCC8,
	hsa04978	Mineral absorption	0.002484	MT1A, MT1E, FXYD2, ATP1B2

We input all the quality IDs in DAVID instrument to observe useful comments and KEGG (Kyoto Encyclopedia of Qualities and Genomes) pathway investigation. Some malignancy related pathways are found in the investigation in Table 6. Further, this pathway investigation additionally uncovers some more targets which oversee pancreatic disease, as GnRH signaling pathway, thyroid hormone flagging pathway, p53 flagging pathway, NF-Kappa flagging pathway and so forth. We find a few qualities among KEGG pathways, for example, RALBP1, TP53, TGFB3, CDH3, PLAU, SMAD3, RAD51, KRAS, CHUK, AKT3, CCND1, RASS5, PLCG2, HDAC1, GNA13, SMAD3 and ADCY7, which are additionally in Pancreatic and other malignancy cell lines and in Table 5. Among these genes a few qualities like AKT3, CHUK, PIK3R3, TGFB3, CDH3, PLAU, RAD51 are as of now known as biomarkers for pancreatic disease [37].

- Conclusion

We have finished our analysis utilizing two existing spectral clustering algorithms and afterward we propose our entropy based changed Shi-Malik spectral clustering algorithms. We have additionally analyzed our proposed framework on Pancreatic malignancy quality articulation dataset and approve the obtained arrangements quantitatively. We additionally perform functional annotation analysis in DAVID instruments to get Gene Ontology comments and KEGG pathways to recognize the Biomarkers for Pancreatic malignancy. We discovered some comparable qualities which act as Biomarkers [36]. Aside from those qualities we have found some more genes that are RASS5, GNA13, ADCY7, SMAD3 which may have noteworthy part as Biomarkers for Pancreatic disease.

CHAPTER 4

❖Ensemble Based Classification of Gene Expression

4.1 Preface

4.2 Methodology

4.3 Experimental Framework

4.4 Result & Discussion

4 Ensemble Based Classification of Gene Expression

4.1 Preface

Malignancy arrange is a basic progress in biomarker recognizing verification. Making machine learning systems that viably expect tumor subtypes can help in recognizing potential malady biomarkers. In this scrutinize, we presented outfit grouping approach and contrasted its execution from other portrayal approaches. PDAC microarray-based quality articulation given in Gene Expression Omnibus (GEO) datasets were investigated. After pre-preparing of information, we classified utilizing Bagged Tree Ensemble strategy and contrasted and different classifiers. The general accomplishment rate henceforth gained was normal of 96.48% for five testing datasets. Such a rate is 6– 15% higher than the looking at rates acquired by various existing DT (decision tree), DA (discriminant analysis) and SVM (support vector machines) and NN (nearest neighbor) approaches, construing that the gathering classifier is astoundingly promising and may transform into a critical test for biomarker distinguishing proof. At last, the organic investigation has been done to recognize the regular biomarkers for PDAC.

4.2 Methodology

Crude CEL documents of five microarray-based quality articulation datasets (GSE28735, GSE15471, GSE41368, GSE32676 and GSE71989) containing articulation information from altogether 105 typical pancreatic and 129 PDAC tissue tests were downloaded from the NCBI Quality Expression Omnibus (GEO) (Table 1). Then the raw file data from each microarray are pre-processed in Matlab2017a. We normalized five datasets and used principal component analysis for feature transformation. Next, we applied Ensemble classifier on clustered solution and compared with nearest neighbour classifier, support vector machine (SVM) classifier, decision tree classifier and discriminant classifier.

- Ensemble Classifier

Ensemble techniques are learning calculations that develop a set of classifiers and after that characterize new information focuses by taking a weighted vote of their predictions. The ensemble

classifiers are bagged tree, boosted tree, subspace discriminant, subspace knn and rusboosted tree which incorporate error correcting [38].

An ensemble of classifiers is an arrangement of classifiers whose individual choices are consolidated somehow, commonly by weighted or unweighted voting to arrange new cases. One of the most dynamic territories of research in administered learning has been to think about techniques for building great groups of classifiers. The primary revelation is that troupes are frequently considerably more precise than the person classifiers that make them up.

Bagged tree, additionally called bootstrap amassing, is a machine learning gathering meta-calculation intended to enhance the security and precision of machine learning calculations utilized as a part of factual order and relapse. It likewise diminishes change and abstains from overfitting [39].Bagged tree train a dataset of size n, bagging generates m new training sets , each of size n′, by sampling from D uniformly and with replacement. By examining with substitution, some observations may be repeated in each . If n′=n, then for large n the set is expected to have the fraction (1 - 1/e) of D, the rest being duplicates. This sort of test is known as a bootstrap sample. The m models are fitted utilizing the above m bootstrap tests and consolidated by averaging the yield or voting [40].

- Nearest Neighbour Classifier

One of the most straightforward choice methods that can be utilized for characterization is the nearest neighbour (NN) method. It orders an example in view of the classification of its closest neighbour. At the point when extensive examples are included, it can be demonstrated that this rule has a probability of error which is less than twice the probability of error contrasted with some other decision rule. The nearest neighbour-based classifiers utilize a few or every one of the examples accessible in the preparation set to order a test design. These classifiers basically include finding the closeness between the test design and each pattern in the preparation set.

The nearest neighbour algorithm assigns to a test pattern the class label of its closest neighbour. Let there be n training patterns, $(X_1,\theta_1),(X_2,\theta_2),...(X_n,\theta_n)$ where X_i is of dimension d and θ_i is the class label of the i^{th} pattern. If P is the test pattern, then if

$$d(P, X_k) = \min\{d(P, X_i)\}$$

Where $i = 1,...n$. Pattern P is assigned to the class θ_k associated with X_k.

The k-nearest neighbours calculation (k-NN) is a non-parametric technique utilized for classification and regression [41]. In both cases, the input comprises of the k nearest preparing cases in the component space. The yield relies upon whether k-NN is utilized for grouping or relapse:

In k-NN characterization, the yield is a class participation. An object is grouped by a dominant part vote of its neighbours, with the protest being allocated to the class most normal among its k closest neighbours (k is a positive number, regularly little). If k = 1, then the object is essentially assigned to the class of that solitary closest neighbour. In k-NN regression, the yield is the property estimation for the object. This esteem is the normal of the estimations of its k nearest neighbours.

This basic method is called the kNN algorithm. There are two major design choices to make: the value of k, and the distance function to use. When there are two alternative classes, to avoid ties the most common choice for k is a small odd integer. An implementation of kNN needs a sensible algorithm to break ties; there is no consensus on the best way to do this. When each example is a fixed-length vector of real numbers, the most common distance function is Euclidean distance:

$$d(x, y) = \|x - y\| = \sqrt{((x - y)\cdot(x - y))} = \left(\sum_{i=1}^{m}(x_i - y_i)^2\right)^{1/2}$$

Where x and y are points in $X = R^m$.

- Support Vector Machine Classifier

SVM is a directed machine learning calculation which can be utilized for both classification or relapse challenges. Here, it is utilized as a part of arrangement issues. In this calculation, it plots every datum thing as a point in n-dimensional space (where n is number of features) with the estimation of each component being the estimation of a specific facilitate. At that point, it performs arrangement by finding the hyper-plane that separate the two classes extremely well. In Fig10 Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes.

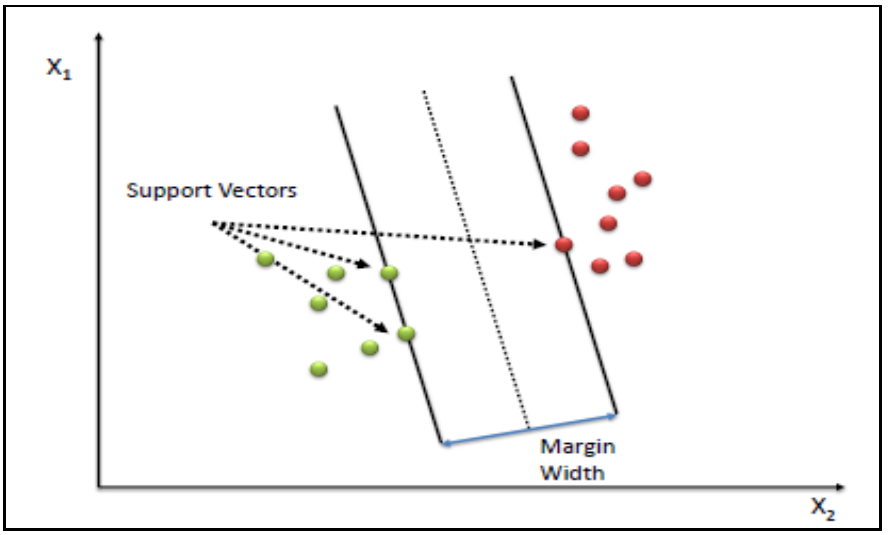


Fig10: Support Vectors

The mappings utilized by SVM plans are intended to ensure that dab items might be processed effectively regarding the factors in the first space, by characterizing them in terms of a kernel function $k(x, y)$ chose to suit the problem [42]. The hyperplanes in the higher-dimensional space are characterized as the arrangement of focuses whose speck item with a vector in that space is consistent. The vectors characterizing the hyperplanes can be been direct mixes with parameters α_i of images of highlight vectors x_i that happen in the information base. With this decision of a hyperplane, the focuses x in the component space that are mapped into the hyperplane are characterized by the connection: $\sum_i \alpha_i k(x_i, x) = cons\tan t$.

If $k(x, y)$ turns out to be little as y becomes advance far from x , each term in the whole measures the level of closeness of the test point x to the relating information base point x_i . Along these lines, the sum of kernels above can be utilized to quantify the relative proximity of each test point to the information focuses starting in either of the sets to be segregated. Note the way that the arrangement of focuses x mapped into any hyperplane can be quite convoluted therefore, permitting significantly more unpredictable segregation between sets which are not arched at all in the first space.

- Decision Tree Classifier

Decision Tree Classifier is a basic and broadly utilized classification procedure. It applies a straightforward thought to tackle the arrangement issue. Decision Tree Classifier represents a progression of painstakingly made inquiries regarding the qualities of the test record. Each time it receives an answer, a subsequent inquiry is asked until a conclusion about the class label of the record is reached [43].

The input features have limited discrete areas, and there is a solitary target include called the classification. Every component of the area of the order is known as a class. A decision tree is a tree in which each inside hub is named with an input feature. The circular segments originating from a hub marked with an info highlight are named with every one of the conceivable estimations of the objective or yield include or the bend prompts a subordinate decision hub on an alternate information include. Each leaf of the tree is named with a class or a probability distribution over the classes. Decision trees can be portrayed additionally as the blend of scientific and computational methods to help the depiction, classification and speculation of a given arrangement of information. Information comes in records of the frame: $(x, Y) = (x_1, x_2, x_3, ...x_k, Y)$

The reliant variable, Y, is the objective variable that we are attempting to comprehend, characterize or sum up. The vector x is composed of the features, x_1, x_2, x_3 and so forth., that are utilized for that assignment.

- Discriminant Analysis Classifier

Discriminant analysis is a factual investigation to foresee a categorical dependent variable by at least one constant or twofold free factors. Discriminant investigation is utilized when bunches are known from the earlier.

Linear discriminant analysis is a speculation of Fisher's straight discriminant, a technique utilized as a part of insights, design acknowledgment and machine figuring out how to locate a direct combination of highlights that describes or isolates at least two classes of articles or occasions. The subsequent combination might be utilized as a direct classifier, or, more commonly, for dimensionality lessening before later characterization. Consider an arrangement of perceptions $\vec{x}$ for each example of a protest or occasion with known class y. This arrangement of tests is known

as the training set. The grouping issue is then to locate a decent indicator for the class y of any example of a similar appropriation given just a perception $\vec{x}$ [44].

LDA approaches the issue by expecting that the contingent likelihood thickness capacities $p(\vec{x}|y=0)$ and $p(\vec{x}|y=1)$ are both regularly circulated with mean and covariance parameters $(\vec{\mu}_0, \Sigma_0)$ and $(\vec{\mu}_1, \Sigma_1)$, separately. Under this presumption, the Bayes ideal arrangement is to anticipate focuses as being from the inferior if the log of the probability proportions is greater than some edge T, so that: $(\vec{x}-\vec{\mu}_0)^T \Sigma_0^{-1}(\vec{x}-\vec{\mu}_0)+\ln|\Sigma_0|-(\vec{x}-\vec{\mu}_1)^T \Sigma_1^{-1}(\vec{x}-\vec{\mu}_1)-\ln|\Sigma_1|>T$ with no further suspicions, the subsequent classifier is alluded to as Quadratic discriminant analysis.

QDA is firmly identified with LDA, where it is expected that the estimations from each class are typically disseminated. In QDA there is no suspicion that the covariance of every one of the classes is indistinguishable. At the point when the ordinariness supposition is valid, the most ideal test for the theory that a given estimation is from a given class is the probability proportion test. Assume there are just two bunches, (so $y \in \{0,1\}$), and the methods for each class are characterized to be $\mu_y = 0, \mu_y = 1$ and the covariances are characterized as $\sum_{y=0}, \sum_{y=1}$. At that point the probability proportion will be given by:

$$\text{Probability Proportion} = \frac{\left(\sqrt{|2\pi\sum_{y=1}|}^{-1} \exp\left(-\frac{1}{2}(x-\mu_{y=1})^T \sum_{y=1}^{-1}(x-\mu_{y=1})\right)\right)}{\left(\sqrt{|2\pi\sum_{y=0}|}^{-1} \exp\left(-\frac{1}{2}(x-\mu_{y=0})^T \sum_{y=0}^{-1}(x-\mu_{y=0})\right)\right)} < t$$

for some edge t. After some revamp, it can be demonstrated that the subsequent isolating surface between the classes is a quadratic. The example evaluations of the mean vector and difference covariance networks will substitute the populace amounts in this equation.

4.3 Experimental Framework

In our experiment, we utilised bagged tree ensemble method on described five PDAC dataset and then compared with nearest neighbour, support vector machine, decision tree and discriminant analysis classifier. Then we analysed the functional enrichment of data using DAVID software and identified the biomarker for PDAC.

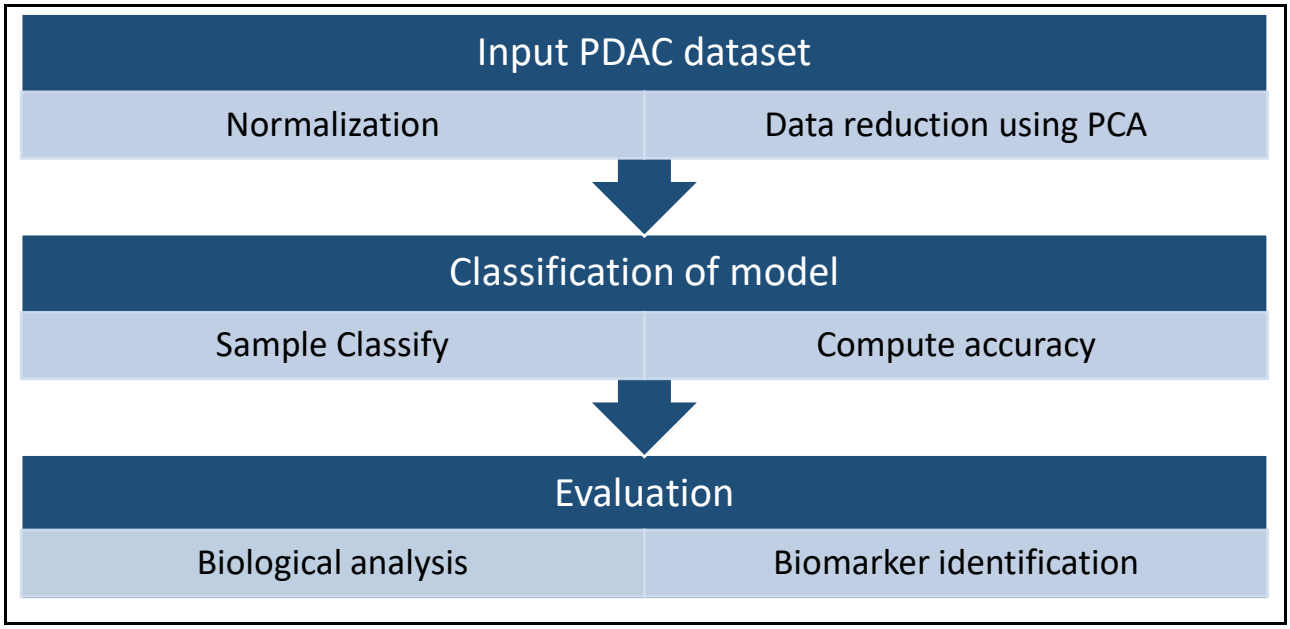


Fig11: Experimental framework of our proposed method

4.4 Result & Discussion

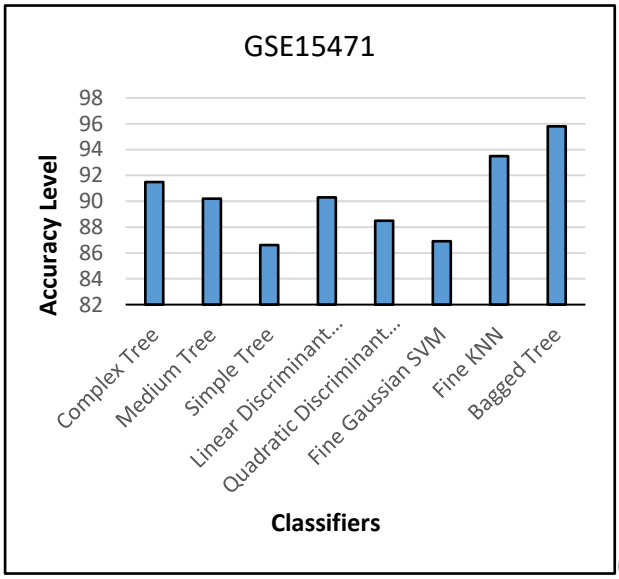
In our work, we normalize five gene expression data set in Matlab2017a using filters and reduce the large gene expression data matrix using principal component analysis. Then we apply classifiers on clustered solution.

we utilize the Complex tree, Medium tree, Simple tree tools in executing the Decision tree classifier. Next, we apply Linear discriminant analysis and Quadratic discriminant analysis toolbox for discriminant analysis classifier, Fine Gaussian SVM toolbox for SVM classifier, Fine KNN toolbox for nearest neighbour classifier and Bagged tree toolbox for ensemble classifier. We receive the 25-fold cross approval precision on the preparation set as the assessment metric and in the comparison with broadly embraced models, we receive the testing precision as the assessment metric. The 25-fold cross validation result of each classifier are shown in Table7. We compare the

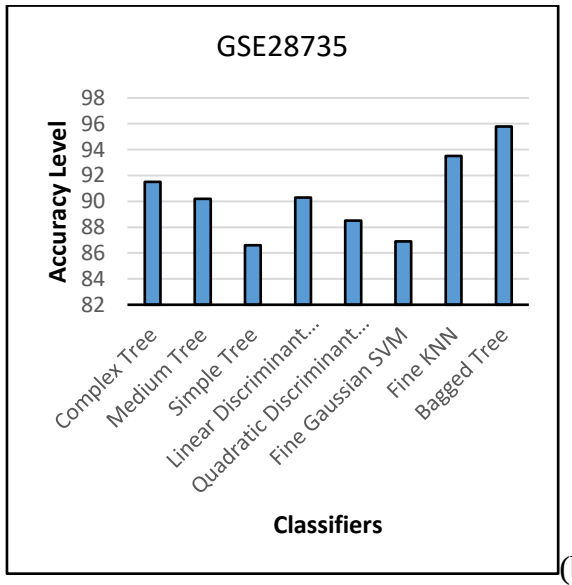
performance of each classifier to obtain best classifier.

Table7: Accuracy level of all classifiers

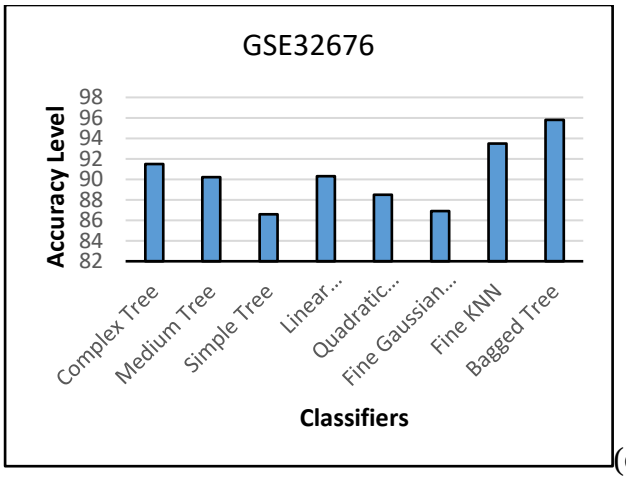
Accuracy Level								
Classifiers	**Decision Tree**			**Discriminant Analysis**		**SVM**	**Nearest Neighbour**	**Ensemble**
	Complex Tree	Medium Tree	Simple Tree	Linear Discriminant Analysis	Quadratic Discriminant Analysis	Fine Gaussian SVM	Fine KNN	Bagged Tree
GSE15471	92.4	91.3	88.1	92.0	83.5	80.9	93.5	**96.8**
GSE28735	91.8	90.7	87.6	94.7	81.7	77.8	93.3	**96.4**
GSE32676	89.6	88.3	81.9	95.3	88.1	82.4	92.1	**96.6**
GSE41368	92.5	90.2	85.4	95.7	92.3	91.5	96.2	**97.0**
GSE71989	91.5	90.2	86.6	90.3	88.5	86.9	93.5	**95.8**



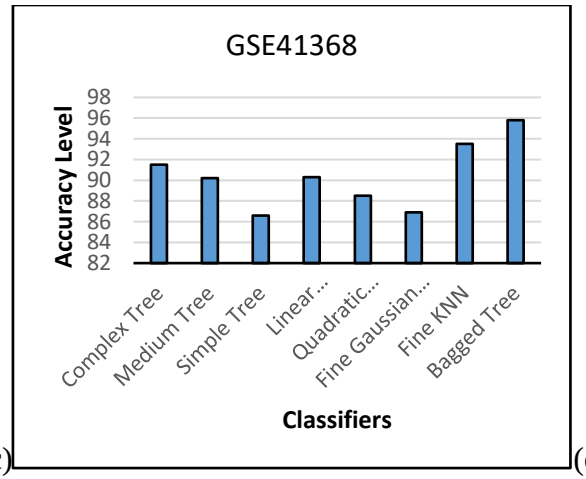
(a)



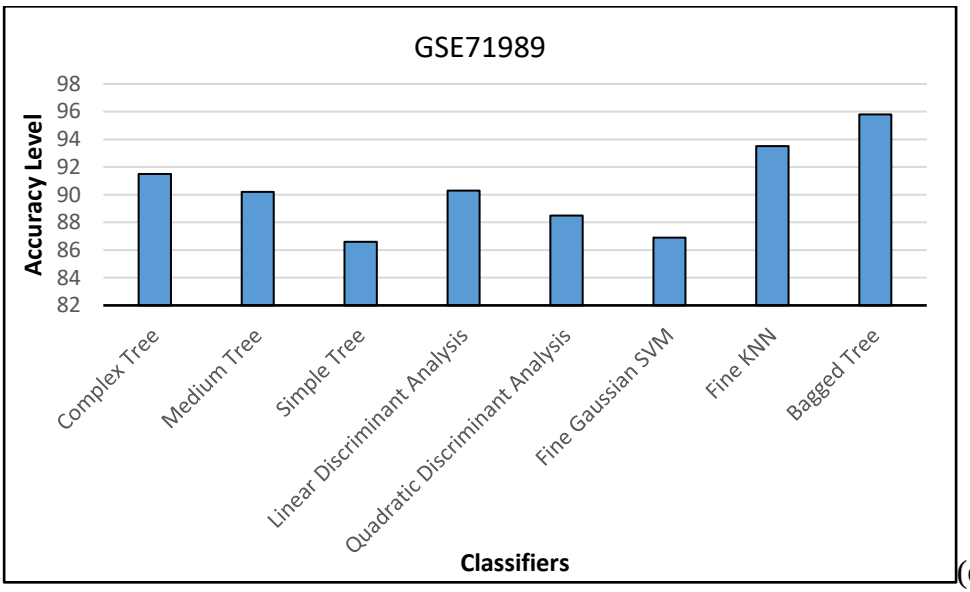
(b)



(c)



(d)



(e)

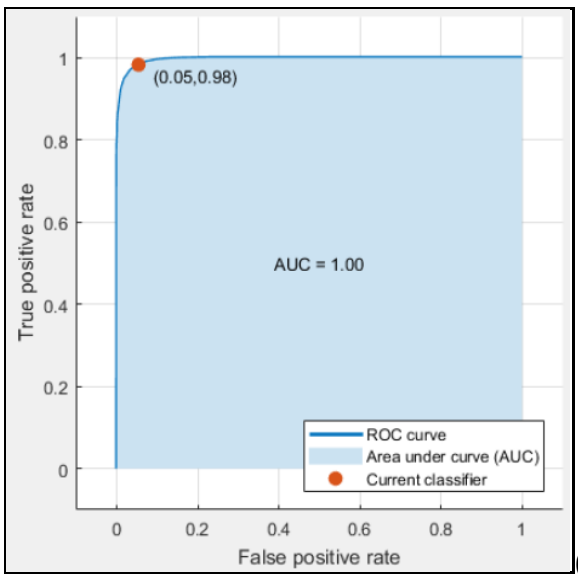
Fig12: 25 fold cross validation accuracy(%) of Ensemble classifier compared to other classifiers for (a) GSE15471, (b) GSE28735, (c) GSE32676, (d) GSE41368 and (e) GSE71989 gene expression.

Figure12 express the accuracy level of different gene expression using five different classifier in which ensemble classifier has the optimal accuracy of average 96.52% against all other classifiers.

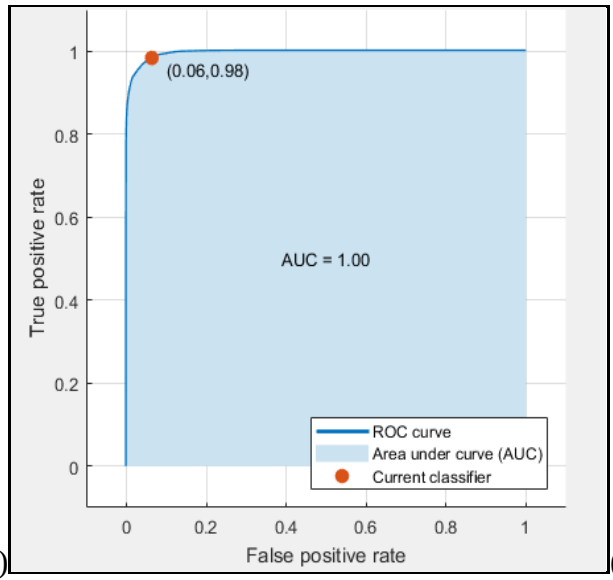
Table8: Validation of Ensemble classifier

Dataset	Ensemble Classifier	
	Specificity	**Sensitivity**
GSE15471	0.9913	0.9968
GSE28735	0.9956	0.9920
GSE32676	0.9968	0.9963
GSE41368	1	1
GSE71989	0.9989	0.9939

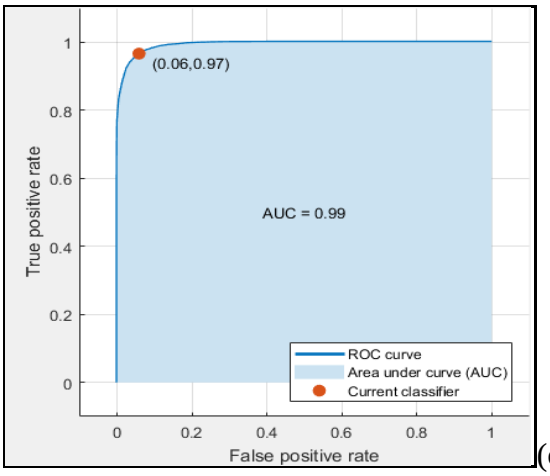
Table8 shows optimal specificity and sensitivity value using ensemble classifier for each dataset.



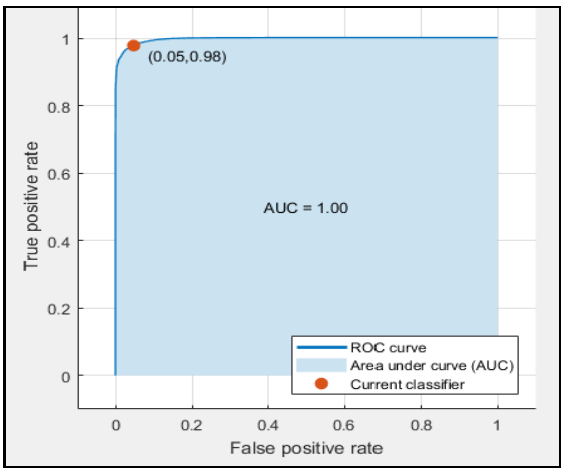
(a)



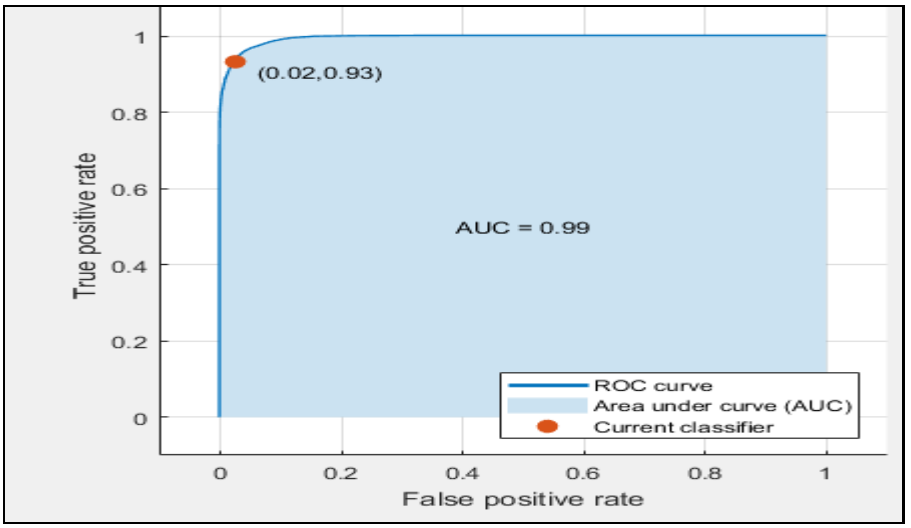
(b)



(c)



(d)



(e)

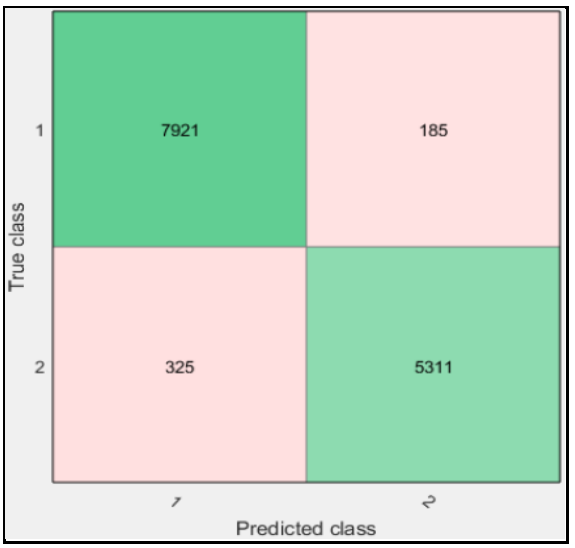
Fig13: ROC of Bagged tree ensemble classifier of five different dataset (a) GSE15471, (b) GSE28735, (c) GSE32676, (d) GSE41368, (e) GSE71989

Fig13 shows the receiver operating characteristic (ROC) curve of selected ensemble classifier for each gene expression dataset which shows true positive rate versus false positive rate.

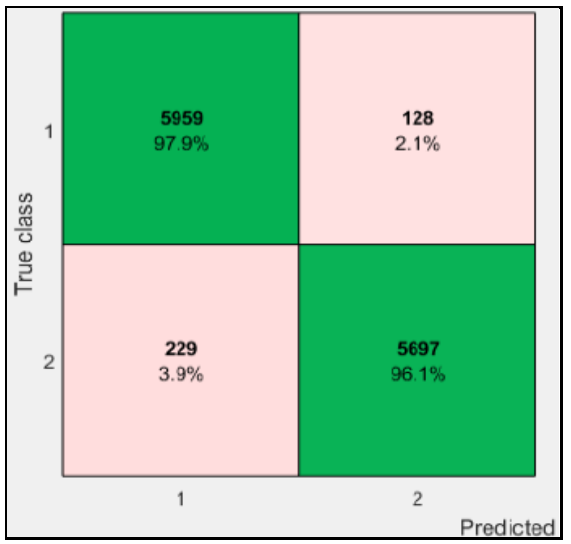
The marker on each plot shows the performance of the selected ensemble classifier. The marker shows the values of the false positive rate (FPR) and the true positive rate (TPR) for the selected classifier. Like in Fig13 (b) false positive rate (FPR) of 0.06 indicates that the current classifier assigns 6% of the observations incorrectly to the positive class whereas a true positive rate of 0.97

indicates that the current classifier assigns 97% of the observations correctly to the positive class. Similarly, we get highest true positive rate for another gene expression using ensemble classifier.

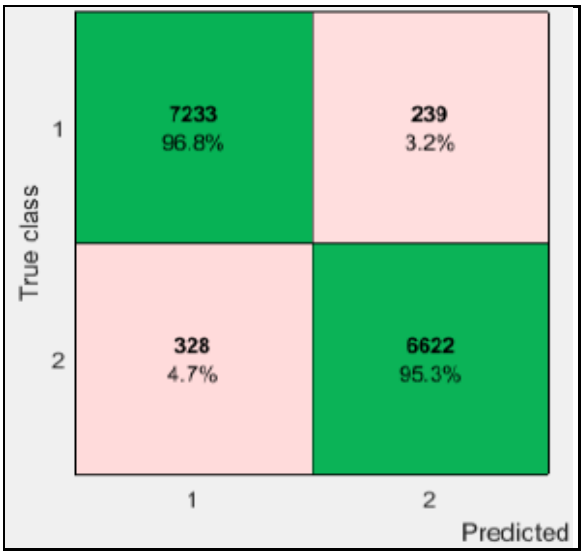
The area under curve is a measure of the overall quality of the classifier. Larger area under curve values indicate better classifier performance. In our work we compare classes and trained models and get best performance in ensemble classifier.



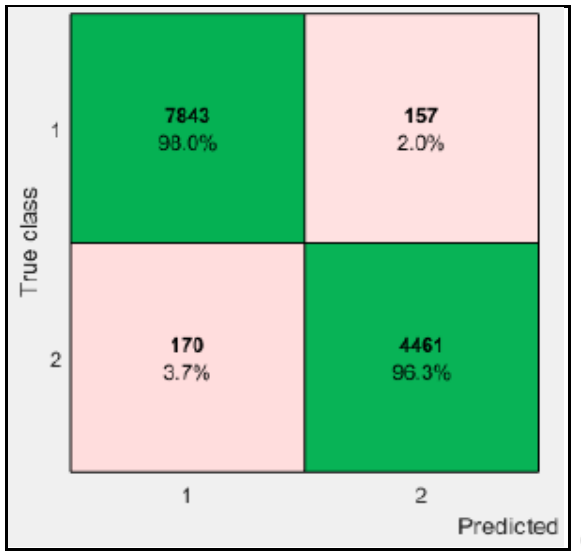
(a)



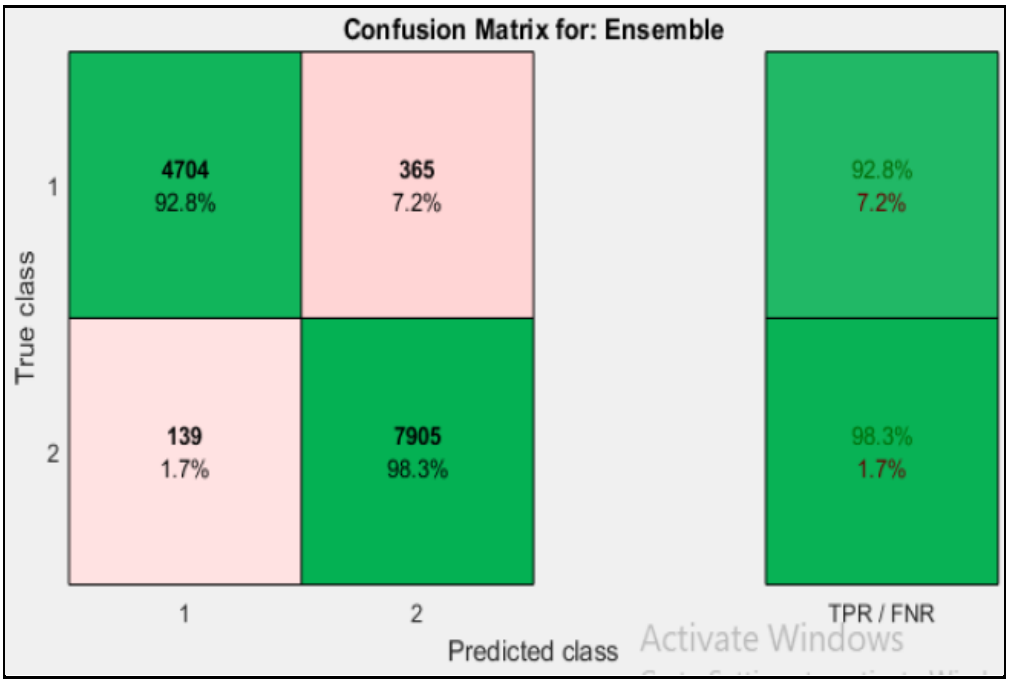
(b)



(c)



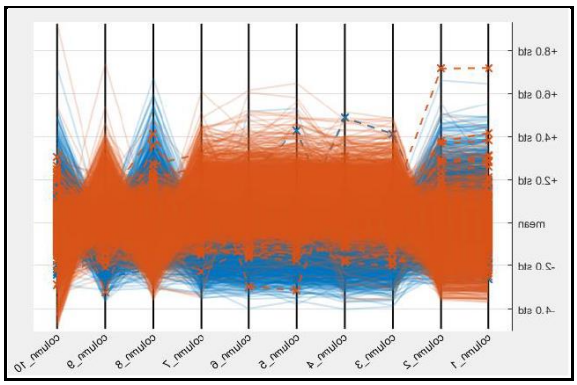
(d)

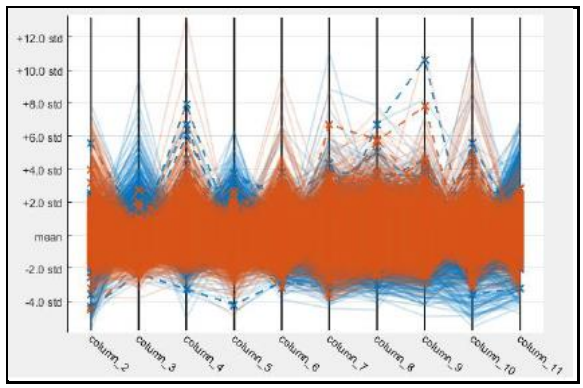


(e)

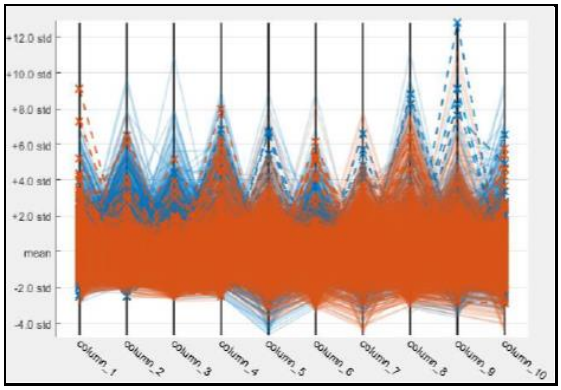
Fig14: CM of Bagged tree ensemble classifier of five different dataset(a) GSE15471, (b) GSE28735, (c) GSE32676, (d) GSE41368, (e) GSE71989

The confusion matrix plot to understand the performance of currently selected ensemble classifier performed in each class. The confusion matrix helps to identify the areas where the classifier has performed poorly. When we open the plot, the rows show the true class, and the columns show the predicted class. The diagonal cells show where the true class and predicted class match. If these cells are green, the classifier has performed well and classified observations of this true class correctly. Fig14 shows the confusion matric (CM) plot of selected ensemble classifier for each gene expression dataset which shows true positive class is much higher than false negative class. The green cells on each plot shows the performance of the selected ensemble classifier Like in Fig14 (b) true class 97.9% of the observations correctly to the positive class. Similarly, we get highest true class for other gene expressions using ensemble classifier.

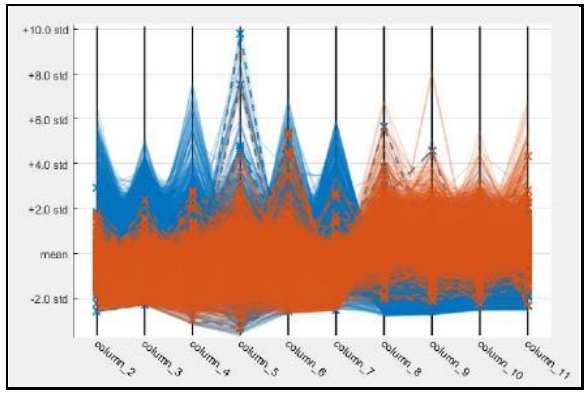

(a)



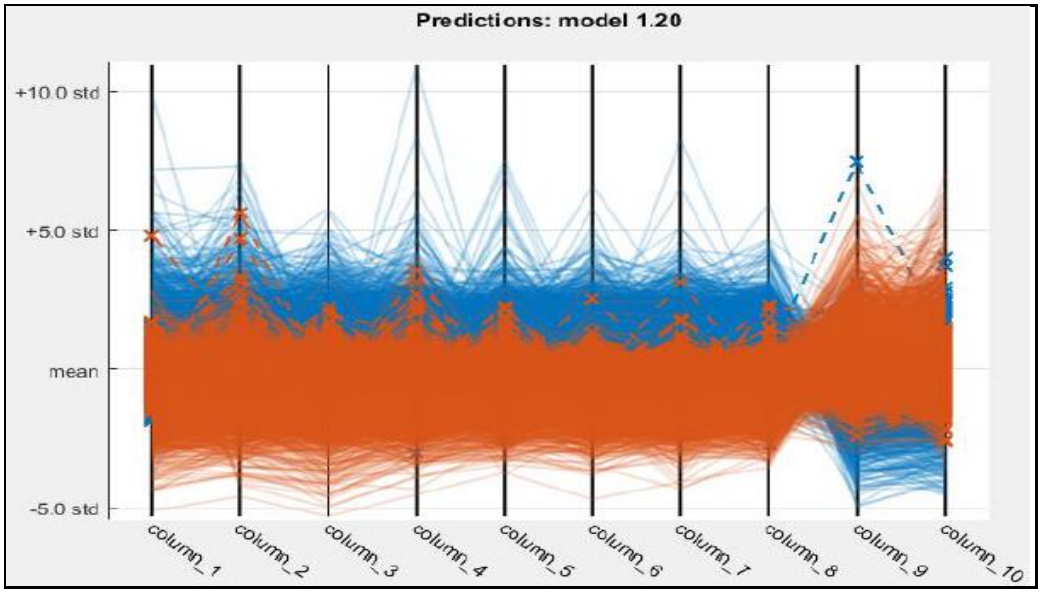
(b)



(c)



(d)



(e)

Fig15: PCP of Bagged tree ensemble classifier of five different dataset(a) GSE15471, (b) GSE28735, (c) GSE32676, (d) GSE41368, (e) GSE71989

We can visualize high dimensional data on a single plot to see 2D patterns. This parallel coordinate plot (PCP) can help us to understand relationships between features and identify useful predictors for separating classes. Fig15 shows training data and misclassified points on the parallel coordinates plot and misclassified points show dashed lines.

- Biological Analysis

We used DAVID apparatus for getting GO annotation. Here we can observe the organic elements of some biomarker qualities. We demonstrate some obtained GO-TERM with p esteem $p<0.05$ and with high % of essence in Table9.

Table9: GOTERM for genes after classification

Gene expression data	**Group**	**GO Term**		**Ontology description**	**Genes**
GSE28735	Cluster1 (Cancer)	BP	GO:0022402	cell cycle process	PRC1, CDC16, SPAG5, RCC2, MAP2K6, NEK9, CLIP1, POLE, DDIT3
		MF	GO:0005515	protein binding	LDHA, RORA, MED21, B2M, MAP3K6, CUL1, SGMS1, MAP3K8
		CC	GO:0005737	cytoplasm	LDHA, REP15, B2M, MAP3K8, CUL5
	Cluster2 (Normal)	BP	GO:0006950	response to stress	A2M, AQP9, S100A9, CUL5, CD48, CFH
		MF	GO:0005515	protein binding	LDHA, RORA, AMOTL1, CUL2, DHX38, CUL5
		CC	GO:0044422	organelle part	B2M, A2M, EPC1, DHX38, HIST2HH2AA4

GSE15471	Cluster1 (Cancer)	BP	GO:0016043	cellular component organization	HIST2H2AA4, EPC1, FAS, TLK2, PTPRM
		MF	GO:0017076	purine nucleotide binding	ADCY4, ADCY7, CCT3, IARS2, MAP3K6, TLK2
		CC	GO:0044446	intracellular organelle part	HIST2H2AA4, A2M, MED21, B2M, DHX38
	Cluster2 (Normal)	BP	GO:0006996	organelle organization	HIS2H2AA4, S100A9, CDC16, TLK2, ROCK1
		MF	GO:0016462	pyro phosphatase activity	GNA13, SEPT4, GNA15, RRAD, RNF213
		CC	GO:0044446	intracellular organelle part	HIS2H2AA4, A2M, REP15, INTS2, MED21
GSE41368	Cluster1 (Cancer)	BP	GO:0042221	response to chemical stimulus	SYT1, ADCY4, A2M, ADCY7, B2M, CD48
		MF	GO:0005524	ATP binding	ADCY4, ADCY7, CCT3, MAP3K6, MAP3K8
		CC	GO:0044422	organelle part	HIST2H2AA4, RAB27B, EPC1, DHX38, A2M
	Cluster2 (Normal)	BP	GO:0002376	immune system process	CADM1, AQP9, STAT5A, B2M, CD48
		MF	GO:0016817	hydrolase activity, acting on acid anhydrides	GNA13, DHX38, GNA15, EIF5, GBP6, DDX24

		CC	GO:0044430	cytoskeletal part	SNGG, SEPT1, VAPA, CDC16, ROCK1, KIF5B
GSE32676	Cluster1 (Cancer)	BP	GO:0006950	response to stress	HIST2H2AA4, EPC1, FAS, TLK2, ROCK1
		MF	GO:0030554	adenyl nucleotide binding	TLK2, MAP2K6, IARAS2, ATP2B4, CCT3
		CC	GO:0044421	extracellular region part	A2M, MMP8, MMP7, CD44, TGFB3
	Cluster2 (Normal)	BP	GO:0000278	mitotic cell cycle	PRC1, CDC16, CUL2, POLE, NCAPD3, RCC2
		MF	GO:0000166	nucleotide binding	ADCY4, ADCY7, CCT3, MAP3K8, TLK2
		CC	GO:0043234	protein complex	PTGS2, INTS2, MED21, EPC1, GNG2, CUL5
GSE71989	Cluster1 (Cancer)	BP	GO:0048518	positive regulation of biological process	S100A6, VAPA, CADM1, STAT5A, STAT5B, CDC16, RORA, B2M
		MF	GO:0001882	nucleoside binding	ROCK1, MARK1, TLK2, MAPK3K8, ADCY7
		CC	GO:0005578	proteinaceous extracellular matrix	LTBP2, POSTN, MMP8, ANG, GPC6, MMP7
	Cluster2 (Normal)	BP	GO:0007049	cell cycle	SEPT4, CDC16, CUL2, TLK2, MAPK3K8

		MF	GO:0032559	adenyl ribonucleotide binding	IARS2, ATP2B4, WNK1, MAP3K8, KIF5B,
		CC	GO:0005737	cytoplasm	ACAA2, KIF5B, LDHA, B2M, CUL5, REP15

We study the biologic and utilitarian relationship of characteristics in our results using GO (Gene Ontology) Annotation gadgets. In Table9 we express some GO Terms for Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) sub ontologies with p regard <0.05. The level of value recommends the characteristics display in GO from all characteristics. The examination of GO Terms exhibit that the batching result is improved with various strategies including cell cycle, cytoplasm, protein complex, nucleotide authoritative, positive control of natural test. Earlier in [45], tantamount GO MF terms, are too uncovered to possibly be associated with Pancreatic threat. We found a couple of characteristics from GO remarks which are generally responsible for malignancy.

Table10: KEGG Pathway analysis for obtained gene expression

Gene expression	**Group**	**KEGG Pathway**	**Pathway name**	**p value**	**Genes**
GSE28735	Cluster1	hsa05200	Pathways in cancer	0.002	FGS7, STAT5A, SPI1, CUL2, FAS
	Cluster2	hsa04510	Focal adhesion	2.16E-04	TLN2, VCL, PDPK1
GSE15471	Cluster1	hsa04912	GnRH signaling pathway	0.002	MAP2K3, KRAS, SOS2, CHUK

	Cluster2	hsa04115	p53 signaling pathway	0.003	CDK1, TP53, IGF1, CHECK1
GSE41368	Cluster1	hsa04062	Chemokine signaling pathway	2.46E-04	ADCY4, STAT5B, CHUK
	Cluster2	hsa04514	Cell adhesion molecules (CAMs)	4.25E-04	ITGAL, CLDN7, CADM1
GSE32676	Cluster1	hsa05211	Renal cell carcinoma	0.003	GRB2, AGNT2, SOS2, AKT3
	Cluster2	hsa03050	Proteasome	0.001	PSMB10, IFNG, PSMB2
GSE71989	Cluster1	hsa05212	Pancreatic cancer	0.01	RALBP1, TP53, SMAD3, CCND1
	Cluster2	hsa04640	Hematopoietic cell lineage	1.92E-04	CSF1, ITGB3, CD9, KITLG

We input all the quality IDs in DAVID gadget to consider handy clarifications and KEGG (Kyoto Encyclopedia of Qualities and Genomes) pathway examination. Some pathways identified with disease are found in the perception in Table 5. Further, this pathway examination furthermore reveals some more targets which are accountable for pancreatic ailment, for example, central bond, GnRH flagging pathway, pancreatic malignancy pathway, renal cell carcinoma, chemokine hailing pathway et cetera. We locate a couple of characteristics among KEGG pathways, for instance, FGS7, STAT5A, SPI1, CUL2, FAS MAP2K3, KRAS, SOS2, ADCY4, ADCY7, GRB2, AGNT2,

AKT3RALBP1, TP53, SMAD3, CCND1, which are in like manner in Pancreatic and other development cell lines and in Table 5. Among these characteristics a couple of characteristics like AKT3, CHUK, KRAS, TP53 are currently known as biomarkers for pancreatic development [46].

- Conclusion

We have completed our order on five PDAC quality articulation dataset using five existing classifiers and got ideal exactness in gathering classifier. We have moreover analyzed our proposed strategy on Pancreatic malignancy quality explanation dataset and favor the procured game plans quantitatively. We similarly perform utilitarian remark examination in DAVID instruments to secure Gene Ontology remarks and what's more KEGG pathways to recognize the Biomarkers for Pancreatic development. We found some tantamount characteristics which go about as Biomarkers [47]. Beside those characteristics we have discovered some more characteristics that are CUL2, FAS, SPI1, GRB2 which may have basic part as Biomarkers for Pancreatic harm.

CHAPTER 5

❖Deep Neural Based Classification of Gene Expression

5.1 Preface

5.2 Methodology

5.3 Experimental Framework

5.4 Qualitative Evaluation

5.5 Functional Enrichment Analysis

5.6 KEGG Pathway Analysis

5.7 Biomarker Identification

5 Deep Neural Based Classification of Gene Expression

5.1 Preface

Threat organize is a fundamental advance in biomarker perceiving confirmation. This abundance of data requires proficient techniques for classification also, examination where deep learning is a promising method for extensive scale information examination. Making deep learning frameworks that suitably expect tumor subtypes can help in perceiving potential disease biomarkers. In this investigate, we introduced furnish gathering approach and differentiated its execution from other depiction approaches. PDAC microarray-based quality explanation given in Gene Expression Omnibus (GEO) datasets were explored. After pre-getting ready of data, we arranged using Stateful Gated Recurrent Unit (GRU) system and differentiated and distinctive classifiers. The general achievement rate from this time forward picked up was typical of 99.22% for five testing datasets. Such a rate is much higher than the taking a gander at rates procured by different existing Stateful Long Short-Term Memory (LSTM), Stateful Recurrent Neural Network (RNN), Stateless GRU, Stateless LSTM and Stateless RNN approaches, understanding that the social occasion classifier is astoundingly encouraging and may change into a basic test for biomarker recognizing confirmation. Finally, the natural examination has been done to perceive the customary biomarkers for PDAC.

5.1 Methodology

Previously discussed crude CEL documents of five microarray-based quality articulation datasets (GSE15471, GSE28735 GSE32676, GSE41368 and GSE71989) containing articulation information from altogether 105 typical pancreatic and 129 PDAC tissue tests were downloaded from the NCBI Quality Expression Omnibus (GEO) (Table 1). Then we applied DNN Classifiers: Stateful GRU, Stateful LSTM, Stateful RNN, Stateless GRU, Stateless LSTM and Stateless RNN approaches on pre-processed dataset.

- Gated Recurrent Unit

GRUs are enhanced variant of standard intermittent neural network. To take care of the vanishing angle issue of a standard RNN, GRU utilizes, purported, update gate and reset entryway. Fundamentally, these are two vectors which choose what data ought to be passed to the yield. GRUs have been appeared to display better execution on littler datasets [48].

There are a few minor departures from the full gated unit, with gating done utilizing the past concealed state and the inclination in different combination, and an improved shape called insignificant gated unit. The fully gated unit is expressed as follows:

It begins with ascertaining the Update Gate z_t for time step t utilizing the formula:

$z_t = \sigma_g \left(W_z x_t + U_z h_{t-1} + b_z \right)$ initially, for $t = 0$, output vector is $h_0 = 0$.

At the point when x_t is connected to the system unit, it is multiplied by its own weight W_z. The same goes for h_{t-1} which holds the data for the past t-1 units and is increased by its own weight U_z. The two outcomes are included and a sigmoid enactment σ_g work is connected to squash the outcome between 0 and 1. A sigmoid capacity is a limited, differentiable, genuine capacity that is characterized for all genuine info esteems and has a non-negative subsidiary at each point. It is a scientific capacity having a trademark S- formed bend or sigmoid curve [49], expressed

as: $S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x + 1}$

The refresh entryway causes the model to decide the amount of the past data should be passed along to the future.

Next Reset Gate r_t is utilized from the model to choose the amount of the past data to overlook.

To figure it, we utilize: $r_t = \sigma_g \left(W_r x_t + U_r h_{t-1} + b_r \right)$

This equation is the same as the one for the update gate. The distinction comes in the weights and the entryway's use, which will find in a bit.

Current Memory content show how precisely the entryways will influence the last yield. In the first place, we begin with the utilization of the update gate. It presents another memory content which will utilize the reset door to store the pertinent data from the past. It is computed as takes

after: $h_t^{'} = \tanh(W x_t + r_t \odot U h_{t-1})$

It increases the information x_t with a weight W and h_{t-1} with a weight U. Then figure the Hadamard item between the reset gate r_t and Uh_{t-1}. That will figure out what to expel from the past time steps. Aggregate up the consequences of previous stages and apply the nonlinear actuation work tanh.

Last Memory at current time step as a last advance, the system needs to compute h_t vector which holds data for the present unit and passes it down to the system. With a specific end goal to do that the update gate is required. It figures out what to gather from the present memory content $h_t^{'}$ and what from the past steps h_{t-1}.

That is done as: $h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h_t^{'}$

where it applied the component-wise multiplication to the update gate z_t and h_{t-1} then multiplied to $(1 - z_t)$ and $h_t^{'}$ finally sum the results from both stage.

GRU able to store and filter information using their update and reset gates. It is deliberately prepared, it can perform amazingly well even in complex situations.

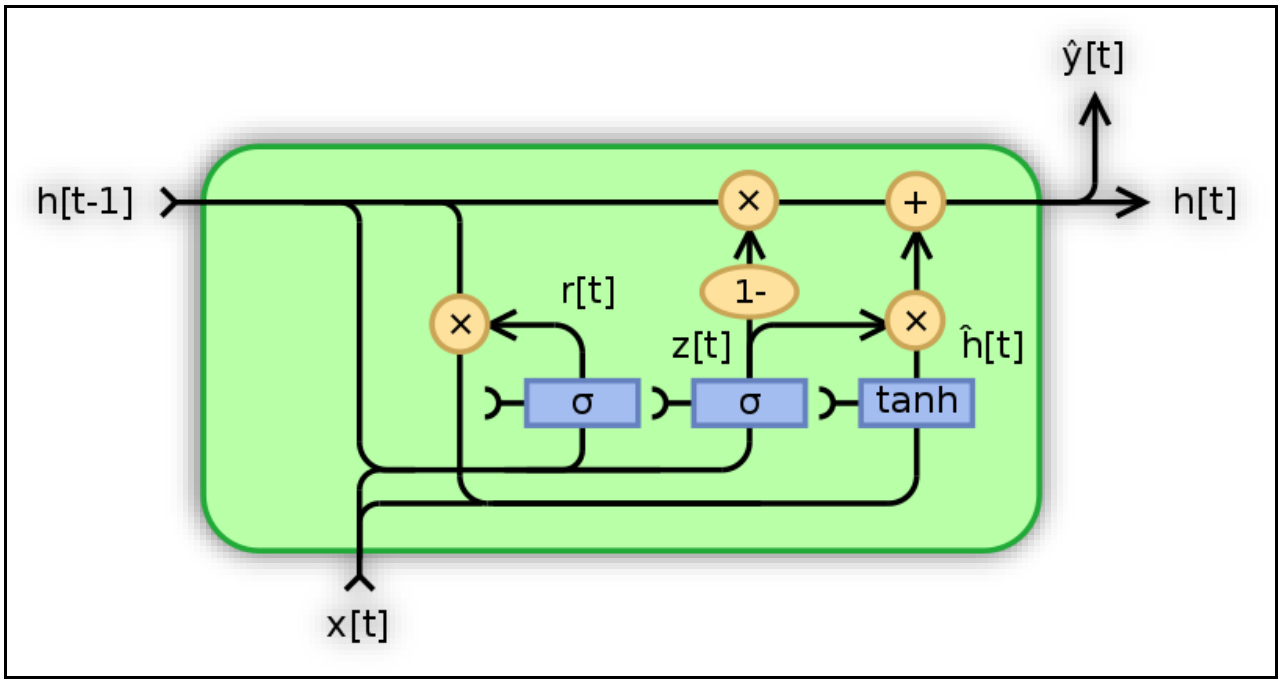


Fig16: Architecture of Fully Gated recurrent Unit

- Long Short-Term Memory

LSTM units are a building unit for layers of RNN [50]. A typical LSTM unit is composed of a cell, an input gate, an output gate and a forget gate [51]. The cell oversees recalling values over discretionary time interims. The articulation long short-term alludes to the way that LSTM is a model for the transient memory which can keep going for a drawn out stretch of time.

The initial phase in LSTM is to choose what data will discard from the cell state. This choice is made by a sigmoid layer called the forget gate layer f_t expressed as follows:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right)$$

It takes a gander at h_{t-1} and x_t yields a number in between 0 and 1 for each number in the cell state C_{t-1}. A 1 speaks to totally keep this while a 0 speaks to totally dispose of this.

The subsequent stage is to choose what new data will store in the cell state. This has two sections. Initial, a sigmoid layer called the input gate layer i_t chooses which esteems will refresh. Next, a $\tanh$ layer makes a vector of new hopeful qualities, $\tilde{C}_t$, that could be added to the state. In the following stage, will join these two to make a refresh to the state as follows:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right)$$
$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right)$$

It's presently time to refresh the old cell state, C_{t-1}, into the new cell state C_t. The past advances effectively chose what to do, simply need to implement it. Then multiply the old state by f_t, overlooking the things we chose to overlook before. At that point it includes $i_t * \tilde{C}_t$. This is the new applicant esteems, scaled by the amount chose to refresh each state esteem.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

At last, it chooses what will yield. This yield will be founded on cell state yet will be a sifted form. In the first place, it run a sigmoid layer which chooses what parts of the cell state will yield. At that point, it put the cell state through tanh to push the qualities to be amongst −1 and 1 and multiply it by the yield of the sigmoid function, with the goal that it just yields the parts it chose to.

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh(C_t)$$

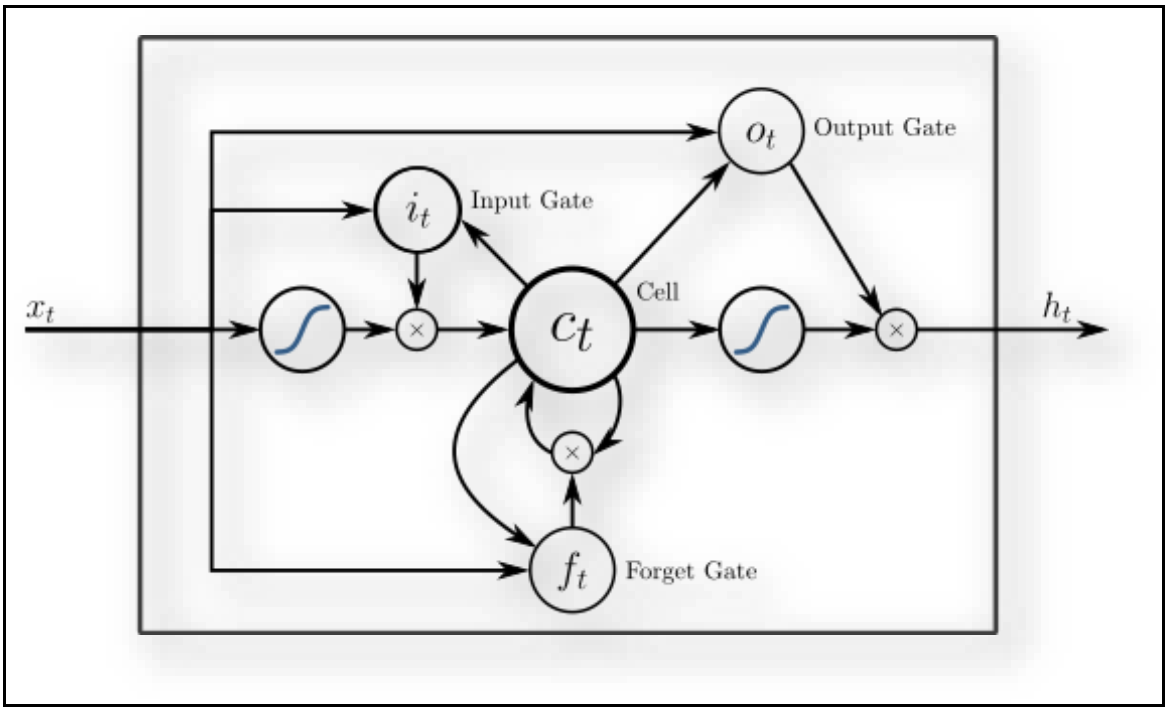


Fig17: Architecture of LSTM system

- Recurrent Neural Network

RNN is a class of counterfeit neural system where associations between units shape a coordinated diagram along an arrangement [52]. This enables it to show dynamic transient conduct for a period arrangement. It can recall sequential incidents and can model time dependencies. It has wide application where the yield depends on previous information. It shares same weights through out every step involved in neural network.

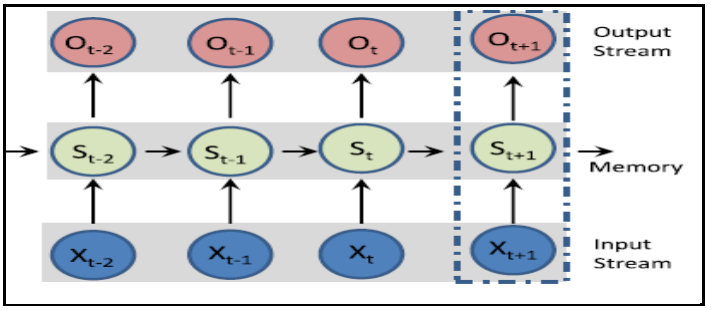


Fig18: Architecture of RNN

In RNN, every one of the sources of info are identified with each other. In request to accomplish it, the RNN makes the systems with circles in them, which enables it to persevere the data.

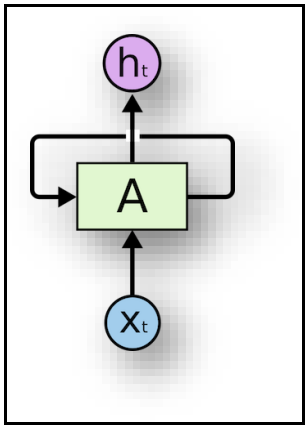


Fig19: RNN step1

This circle structure enables the neural system to take the succession of information. It shows unrolled form in Fig20.

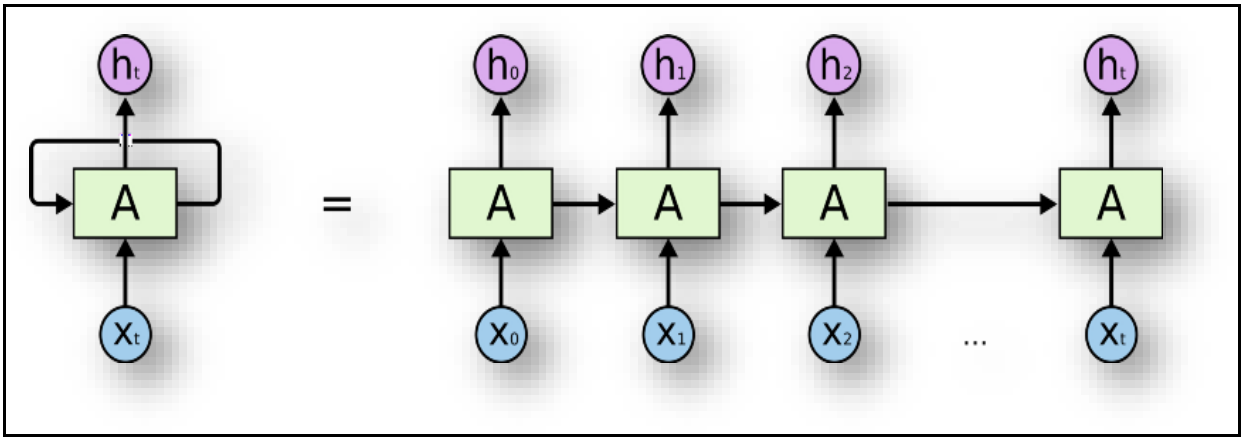


Fig20: Unrolled form of 1st step of RNN

It takes the X_0 from the arrangement of information and after that it yields h_0 which together with X_1 is the contribution for the subsequent stage. Thus, the h_0 and X_1 is the contribution for the following stage. Likewise, h_1 from the following is the contribution with X_2 for the subsequent stage et cetera. Along these lines, it continues recollecting the unique circumstance while preparing. Thus, recurrent neural network works and yield best output using computation data.

Here, we used both stateless and stateful GRU, LSTM and RNN for classification of our dataset. A stateless framework can be viewed as a container where anytime the estimation of the output depends just on the estimation of the input after a specific handling time.

A stateful framework rather can be viewed as a crate where anytime the estimation of the output relies upon the estimation of the input and of an inside state, so basically a stateful framework resembles a state machine with memory as a similar arrangement of input esteem can produce distinctive output contingent upon the past input got by the framework.

5.3 Experimental Framework

In our experiment, we utilized stateful and stateless recurrent neural network methods on described five PDAC dataset and then compared with GRU, LSTM and simple RNN. Then we analyzed the functional enrichment of data using DAVID software and identified the biomarker for PDAC.

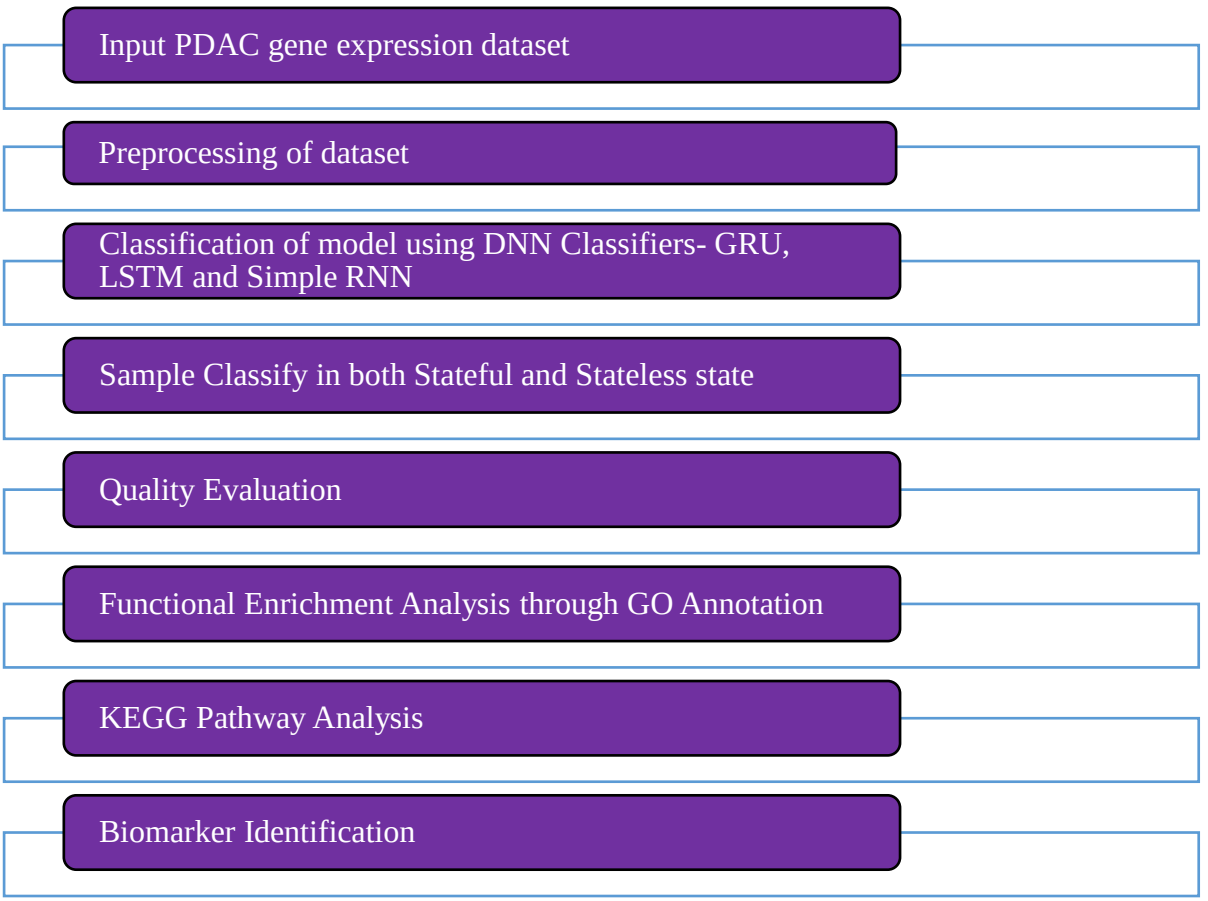


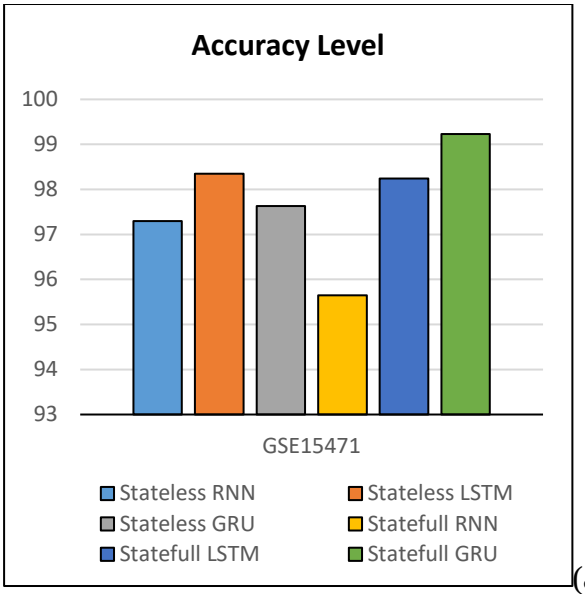
Fig21: Experimental framework of proposed Deep learning method

5.4 Quality Evaluation

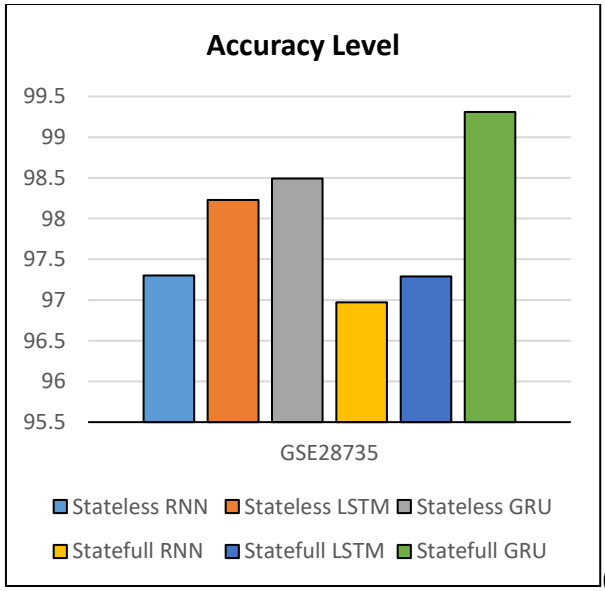
We utilized stateful, stateless neural network classifiers which classified dataset in Python using batch size 1 and epoch 32 with activation layer sigmoid function and tanh. An epoch is a measure of the number of times all the training vectors are used once to update the weights. For batch training all the training samples pass through the learning algorithm simultaneously in one epoch before weights are updated.

Table11: Accuracy level of all neural classifiers

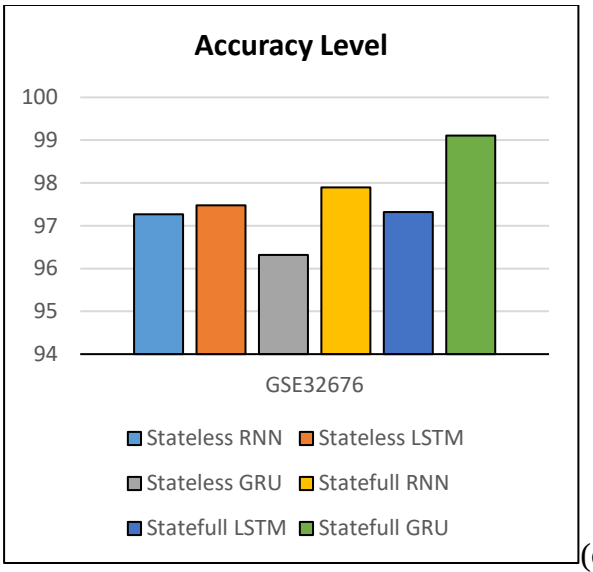
Accuracy Level						
Classifiers	**Stateless Classifier**			**Stateful Classifier**		
	RNN	**LSTM**	**GRU**	**RNN**	**LSTM**	**GRU**
GSE15471	97.30	98.35	97.63	95.65	98.24	**99.23**
GSE28735	97.30	98.23	98.49	96.97	97.29	**99.31**
GSE32676	97.27	97.48	96.32	97.90	97.32	**99.11**
GSE41368	96.34	98.56	96.10	96.94	95.56	**99.40**
GSE71989	94.51	96.48	97.11	98.84	96.71	**99.02**



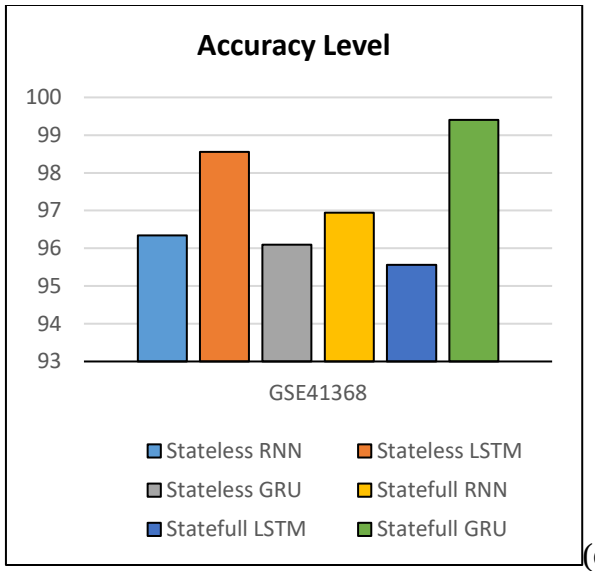
(a)



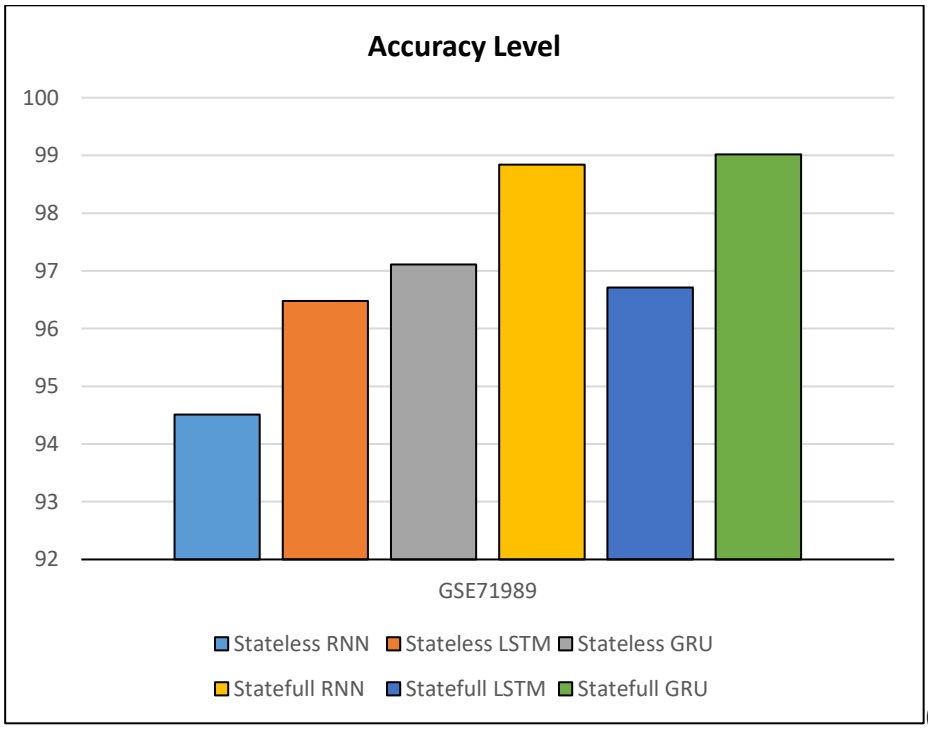
(b)



(c)



(d)



(e)

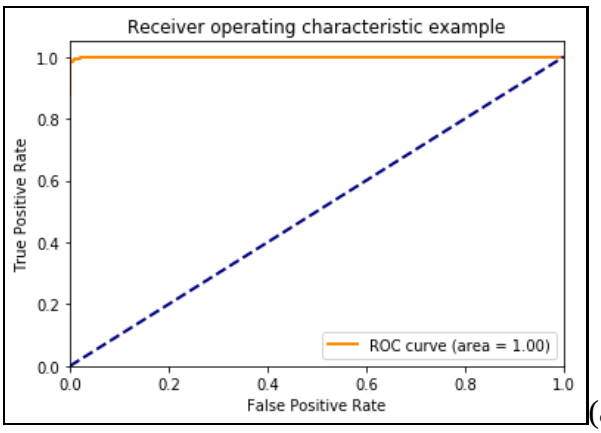
Fig21: Accuracy(%) of Neural Network classifier compared to other classifiers for (a) GSE15471, (b) GSE28735, (c) GSE32676, (d) GSE41368 and (e) GSE71989 gene expression.

Figure21 express the accuracy level of different gene expression using three different classifier in both statefull and stateless state in which stateful GRU has the optimal accuracy of average 99.22% against all other neural classifiers.

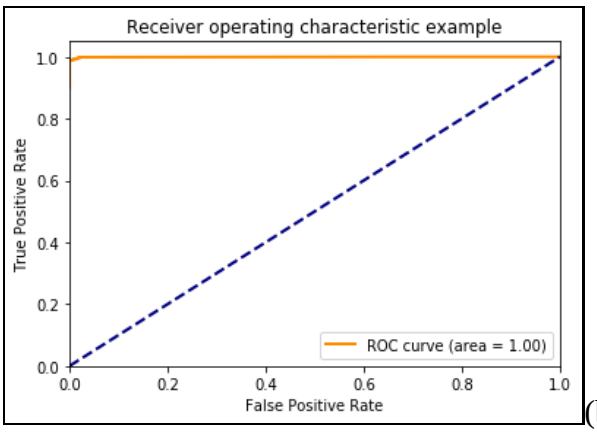
Table12: Validation of Neural classifier

Dataset	Statefull Gated Recurrent Unit	
	Specificity	**Sensitivity**
GSE15471	0.9901	0.9928
GSE28735	0.9891	0.9937
GSE32676	0.9915	0.9952
GSE41368	0.9973	0.9945
GSE71989	0.9849	0.9901

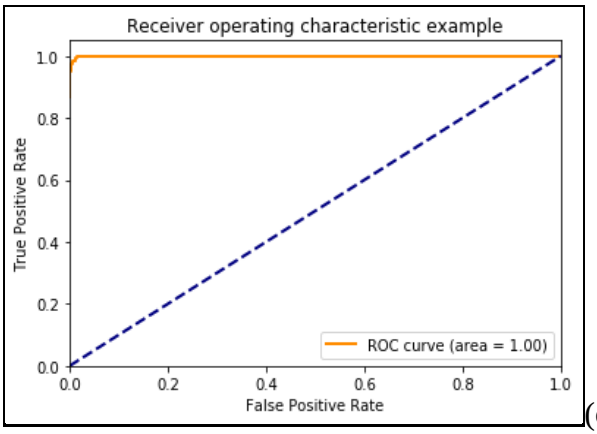
Table12 shows optimal specificity and sensitivity value using statefull GRU classifier for each dataset.



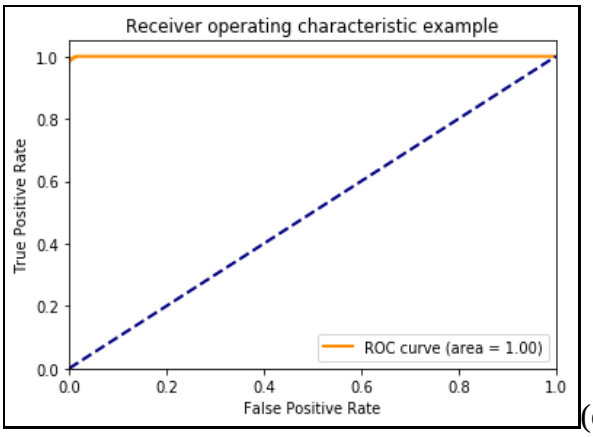
(a)



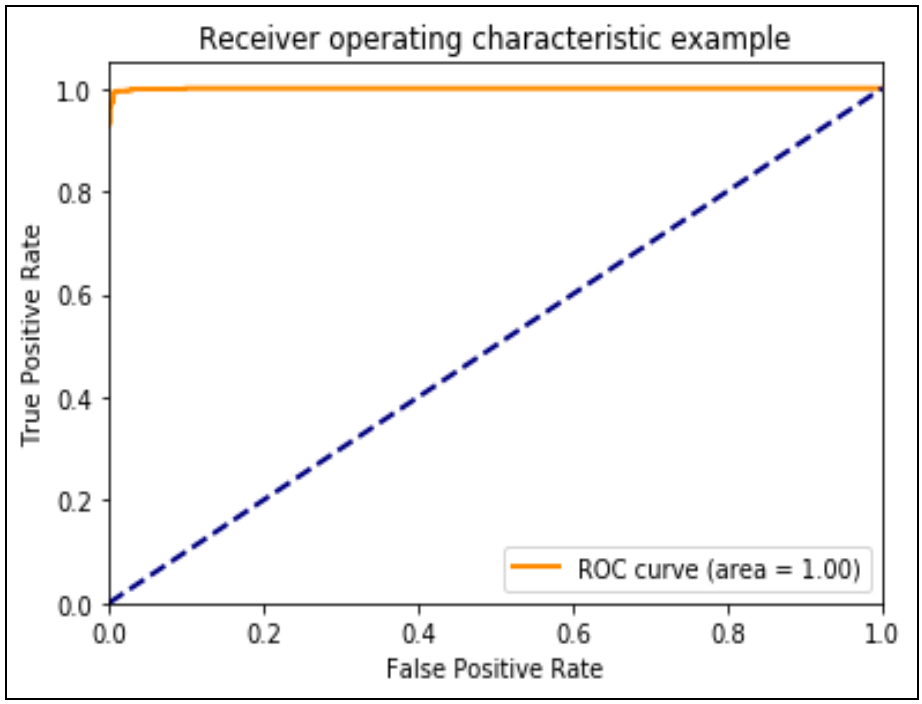
(b)



(c)



(d)



(e)

Fig23: ROC of Stateful GRU neural classifier of five different dataset (a) GSE15471, (b) GSE28735, (c) GSE32676, (d) GSE41368, (e) GSE71989

Fig23 shows the receiver operating characteristic (ROC) curve of selected stateful GRU neural classifier for each gene expression dataset which shows true positive rate versus false positive rate.

The area under curve is a measure of the overall quality of the classifier. Larger area under curve values indicate better classifier performance. In our work we compare classes and trained models and get best performance in stated classifier.

5.5 Functional Enrichment Analysis

We observed functional enrichment analysis through GO annotation using DAVID tool software as described in Table13.

Table13: GO Annotation for gene expression

Gene expression data	Group	GO Term		Ontology description	Genes
GSE15471	Cluster1 (Cancer)	BP	GO:0010033	response to organic substance	ADCY4, A2M, APOBEC1, THRA, PTGS2, ADCY7, AQP9, STAT5A, ARNT2, STAT5B, TGFB3, PMAIP1, MMP3, IL10, TGFB2, B2M, CD48, CD44, GATA3, SMAD3, MALT1, SMAD2.
		MF	GO:0001883	purine nucleoside binding	ADCY4, ADCY7, IARS2, CCT3, PRKG1,ATP2B1,MAP3K6,ATP2B4, DHX38,MAP3K8,DYNC2H1,TLK2, ROCK1P1, SLFN12, KIF20B, AKT3.
		CC	GO:0000267	cell fraction	CADM1, VAPA, AQP9, PTGS2, C16ORF70, LEMD3, CD52, CUL5, FAS, SLC12A6, PTPRF, CRYAB, GRIN2A, WNK1, TPI1P1, JUP, VSIG2, PTRF, NME1, LAMC2, TMX1, GCNT3, GNAI3,CHUK.
	Cluster2 (Normal)	BP	GO:0051301	cell division	KIF23, SEPT4, PRC1, SEPT1, CDC14A, TSG101, KNTC1, CDC16, LATS2, TGFB2, CCNE1, NDE1, SEH1L, NUP37, CABLES1, ASPM, CDC7, CDK1, CDC6, ROCK1.
		MF	GO:0003924	GTPase activity	RHOJ, GNA13, SEPT4, GNA15, ATL1, ATL3, EIF5, RRAD, RHOU, ARHGAP5,RHOC,TUBG1,RAB27B , GNL2, TUBA1A, RHOF
		CC	GO:0044444	cytoplasmic part	LDHA, A2M, PTGS2, REP15, C16ORF70, SGMS1, B2M, CUL5, BTBD1, CH25H, MAP3K8, PHTF1, GNG2, FAS, RAB27B,AKT3.

GSE28735	Cluster1 (Cancer)	BP	GO:0032787	monocarboxylic acid metabolic process	SCPEP1, PTGES3, LDHA, CYP2J2, PTGS2,STAT5A,CRABP2,STAT5B, NDUFAB1,ACOT1,BBOX1,CHUK.
		MF	GO:0017076	purine nucleotide binding	ADCY4, ADCY7, CCT3, IARS2, PRKG1,ATP2B1,MAP3K6,ATP2B4, DHX38, MAP3K8, DYNC2H1, RAB25, TLK2, RAB27B.
		CC	GO:0044446	intracellular organelle part	HIST2H2AA4, A2M, PTGS2, REP15, C16ORF70, INTS2, SGMS1, MED21, ANKLE2, B2M, EPC1, DHX38, INTS7, PHTF1, RAB27B, RAB27A, ACAA2, PDPR, NCF2, ROCK1,
	Cluster2 (Normal)	BP	GO:0009725	response to hormone stimulus	ADCY4, A2M, APOBEC1, THRA, LDLR, ADCY7, PTGS2, STAT5A, LEPR, STAT5B, ARNT2, TGFB3, IL10, LATS2, TGFB2, CCNE1, PDPK1, ANG
		MF	GO:0004175	endopeptidase activity	PDIA3, LGMN, MMP8, MMP7, MIPEP, MMP3, MMP2, MMP1, CMA1, RHBDL2, PRSS23, PLAU
		CC	GO:0009986	cell surface	CSPG4, TGFB3, CD53, CD48, ACVR1B, NOD2, S1PR1, CD44, GPC6, MS4A2, CEP290, FAS, KLRD1, SPN, ICAM1, STX4, CLCA1, ADAM10, CD3E, CRYAB, FLOT2, TGFBR3, CLEC7A.
GSE32676	Cluster1 (Cancer)	BP	GO:0042221	response to chemical stimulus	SYT1, ADCY4, A2M, PTGS2, ADCY7, AQP9, STAT5A, S100A9, STAT5B, PDLIM1, B2M, CD48, PLOD1, CD44, CCR10, IFNG, CHRNA5, GNG2, FAS, HTR1D, MAP2K6, , MMP12, GNAL.
		MF	GO:0070011	peptidase activity, acting on L-amino acid peptides	SCPEP1, CNDP2, LGMN, MMP8, MMP7, MIPEP, MMP3, MMP2, MMP1, HTRA1, ADAM8, OMA1, CHUK, SMAD3.

		CC	GO:0043005	neuron projection	ADCY4, SNCG, SYT1, NRP1, CADM1, ATL1, C16ORF70, IQGAP1, TGFB2, ANG, HDC, CHRNA5, CNTNAP1, ATP6V0D1, NEGR1, SAMD4A, RAB27A.
	Cluster2 (Normal)	BP	GO:0042981	regulation of apoptosis	CADM1, PTGS2, STAT5A, ARNT2, STAT5B, TGFB3, PMAIP1, SGMS1, TNFSF18, IL10, TGFB2, CUL2, G2E3, CUL5, CD44, IFNG, FAS, NQO1, API5, MAP2K6, SPN, NET1, RAB27A, CD3G, SOCS2, PTPRF, ROCK1, CD3E, ID3, APAF1.
		MF	GO:0017111	nucleoside-triphosphatase activity	GNA13, SEPT4, GNA15, EIF5, RRAD, RNF213, ATP2B1, ATP2B4, DHX38, STARD9,DDX24, GTPBP4, DYNC2H1, DDX21, TUBG1, RAB27B, GNL2, RAB27A,
		CC	GO:0031012	extracellular matrix	LTBP2, MMP8, MMP7, TGFB3, POSTN, MMP3, MMP2, MMP1, MMRN2, TGFB2, HMCN1, CD44, ANG, GPC6, SERPINA1, COL11A1, LOXL1, SPN, SPON1, STX4, LAD1.
GSE41368	Cluster1 (Cancer)	BP	GO:0006996	organelle organization	HIST2H2AA4, S100A9, CDC16, PRKG1, EPC1, DYNC2H1, NUP37, TLK2, ROCK1, CRYAB, RCOR1, OPTN, NCAPD3, MARK1, DCTN2, NCAPD2, THY1, KRT19, RND1, RCC2, RHOF, CDK1, , KDM4A, RNF40, ARAP1.
		MF	GO:0031701	angiotensin receptor binding	GNA13, EDNRB, GNB1, AGT, BDKRB2,TP53,CDK,CHUK.
		CC	GO:0032991	macromolecular complex	HIST2H2AA4, PTGS2, INTS2, MED21, B2M, SPRY2, NDE1, GTF2A2, MRPL17, BCL2, ABCD2, THBS1, ZWILCH, AGL, ETV3, BCAS2, CD1C, CD1B,

	Cluster2 (Normal)	BP	GO:0009987	cellular process	GNA13, SYT1, A2M, AP4E1, CADM1, C16ORF70, TGFB3, TGFB2, DMXL2, SLC2A1, RAB27B, RAB21, RAB27A, CLCA1,
		MF	GO:0005509	calcium ion binding	SYT1, S100A6, LTBP2, S100A8, LTBP3, STAT5A, STAT5B, S100A9, NELL2, MMP8, NDUFAB1, MMP7, MIPEP, DUOX2, ANO1.
		CC	GO:0043234	protein complex	GNA13, SYT1, A2M, AP4E1, CADM1, C16ORF70, TGFB3, TGFB2, DMXL2, SLC2A1, RAB27B, RAB21, RAB27A, CLCA1.
GSE71989	Cluster1 (Cancer)	BP	GO:0031401	positive regulation of protein modification process	TGFB3, CDC16, PSMD4, CDK1, CD3E, IL6R, PSMA1, CCND1, PSEN1, PSMA6, CCND2, PIAS3, PSME2, PSMA5, PSMA4, CD81, PSMA3, UBC, PIAS1, LRRK2.
		MF	GO:0046983	protein dimerization activity	S100A6, THRA, VAPA, CADM1, CNDP2, HEXA, ARNT2, TGFB3, TGFB2, PLOD1, TRIM8, GTF2A2, ELTD1, TOP2A, GABPB2, CARS, CD3G, ADAM10, CD3D, CD3E, TP53, ADIPOR2, IL6R, JUNB, DDIT3, C1QB, HIF1A, SBF2, SDS, CTSE, CLIP1, NFE2L1.
		CC	GO:0045177	apical part of cell	PRKCZ, LGMN, ERBB2, DUOX1, LMO7, CACNB3, VCAM1, AKR1A1, CD44, P2RY2, SCNN1G, ATP6V0D1, SCNN1A, RAB27A, STX4, MYO1A, INADL, CUBN, STX2, CLCA4, PDPN, STXBP3, MYL12B, IL6R, PFKM, SLC9A3R1, P2RX4, NEDD1, ITGA8, SLC26A9, USH1C.
	Cluster2 (Normal)	BP	GO:0010646	regulation of cell	GPR65,FHL3,HK1,EEA1,GNG12,MYCBP2,UEVLD,PDE1,CRISPLD2,ALOX5AP,SOS2,REXO2,HSD17B6,R

				communicati on	AB11A,TRIP10,PCSK6,BUB3,ZNF2 67, MARS, SELP, SWAP70, SELL.
		MF	GO:0032559	adenyl ribonucleotid e binding	IARS2,ATP2B4,WNK1,MAP3K8, KIF5B, CHUK.
		CC	GO:0044446	intracellular organelle part	HIST2H2AA4, A2M, PTGS2, REP15, C16ORF70, INTS2, SGMS1, MED21, ANKLE2, B2M, EPC1, DHX38, INTS7, PHTF1, RAB27B, CLTC, ZZEF1, ALOX5, ACTR10.

We elucubrate the functional enrichment relationship of characteristics in our results using GO (Gene Ontology) Annotation gadgets. In Table 13 we express some GO Terms for Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) sub ontologies with p regard <0.05. The level of value proposes the characteristics display in GO from all characteristics. The examination of GO Terms show that the batching result is improved with various techniques including response to organic substances, purine nucleoside binding, cell fraction, cell division, GTPase activity, cytoplasmic part, monocarboxylic acid metabolic process, intracellular organelle part, response to hormone stimulus, endopeptidase activity. Earlier in [45], equivalent GO MF terms, are too uncovered to be in any way associated with Pancreatic threat. We found a couple of characteristics from GO remarks which are generally responsible for threat.

5.6 KEGG Pathway Analysis

Table14: KEGG Pathway analysis for obtained gene expression

Gene expression	Group	KEGG Pathway	Pathway name	p value	Genes
GSE15471	Cluster1 (Cancer)	hsa05200	Pathways in cancer	0.002	FGF7,PTGS2,FGF14,STAT5A, STAT5B,ARNT2,SPI1,TGFB3, NFKB2,MMP2,MP1,TGFB2, CUL2,ACVR1B,CCNE1, SLC2A1,CSF3R,FAS,TPR, CHUK, AKT3, PRKCA, RALBP1,TP53,CDK2,SMAD4.

	Cluster2 (Normal)	hsa04662	B cell receptor signaling pathway	8.15E-4	PTPN6, VAV3, CR2, IFITM1,GRB2,NFKBIA, MALT1, VAV1, PRKCB, NRAS, KRAS, FCGR2B, SOS2, CD81, PLCG2, MAPK3,PPP3CB,CD79B.
GSE28735	Cluster1 (Cancer)	hsa04115	p53 signaling pathway	0.0003	CDK1, TP53, IGF1, CHEK1, PMAIP1, SFN, CDK2, SESN3, RFWD2,CCNE1,CCND1,CCNB2, CCND2,SERPINB5,CD82,DDB2, MDM2, APAF1, FAS, THBS1, GADD45A.
	Cluster2 (Normal)	hsa04664	Fc epsilon RI signaling pathway	0.0003	PRKCA, FCER1A, VAV3, PLA2G10, GRB2, MAP2K3, VAV1, PRKCB, NRAS, PLA2G4A, KRAS, GAB2,SOS2, PLCG2,MAPK3,PLA2G2A, MS4A2, FCER1G, MAPK8, PIK3R3, MAP2K6, AKT3.
GSE32676	Cluster1 (Cancer)	hsa05220	Chronic myeloid leukemia	0.001	GRB2, STAT5A, STAT5B, CBL, TP53, TGFB3, NFKBIA, SMAD3, TGFB2, ACVR1B, NRAS, CCND1, KRAS, GAB2, HDAC1, SOS2, MAPK3, MDM2, PIK3R3, CHUK.
	Cluster2 (Normal)	hsa04960	Aldosterone-regulated sodium reabsorption	0.0001	PRKCA, PDPK1, KRAS, MAPK3, HSD11B1, IGF1, SCNN1G, SFN, PIK3R3, SCNN1A, PRKCB.
GSE41368	Cluster1 (Cancer)	hsa05223	Non-small cell lung cancer	0.002	PRKCA, GRB2, ERBB2, TP53, PRKCB, NRAS, RASSF5, CCND1, PDPK1, KRAS, SOS2, PLCG2, MAPK3, PIK3R3, AKT3.
	Cluster2 (Normal)	hsa03050	Proteasome	0.001	PSMB10, IFNG, SMAD3, PSMB2, MAPK4, MAPK3.
GSE71989	Cluster1 (Cancer)	hsa05212	Pancreatic cancer	0.01	RALBP1, TP53, SMAD3, CCND1, SOS2, STAT5.
	Cluster2 (Normal)	hsa04670	Leukocyte trans endothelial migration	4.23E-10	ITGAL, CLDN7, GNAI3, ITGB1, MMP2, ITGAM, PXN, CDH5, VCL, VAV1, CTNNA3, THY1, JAM3.

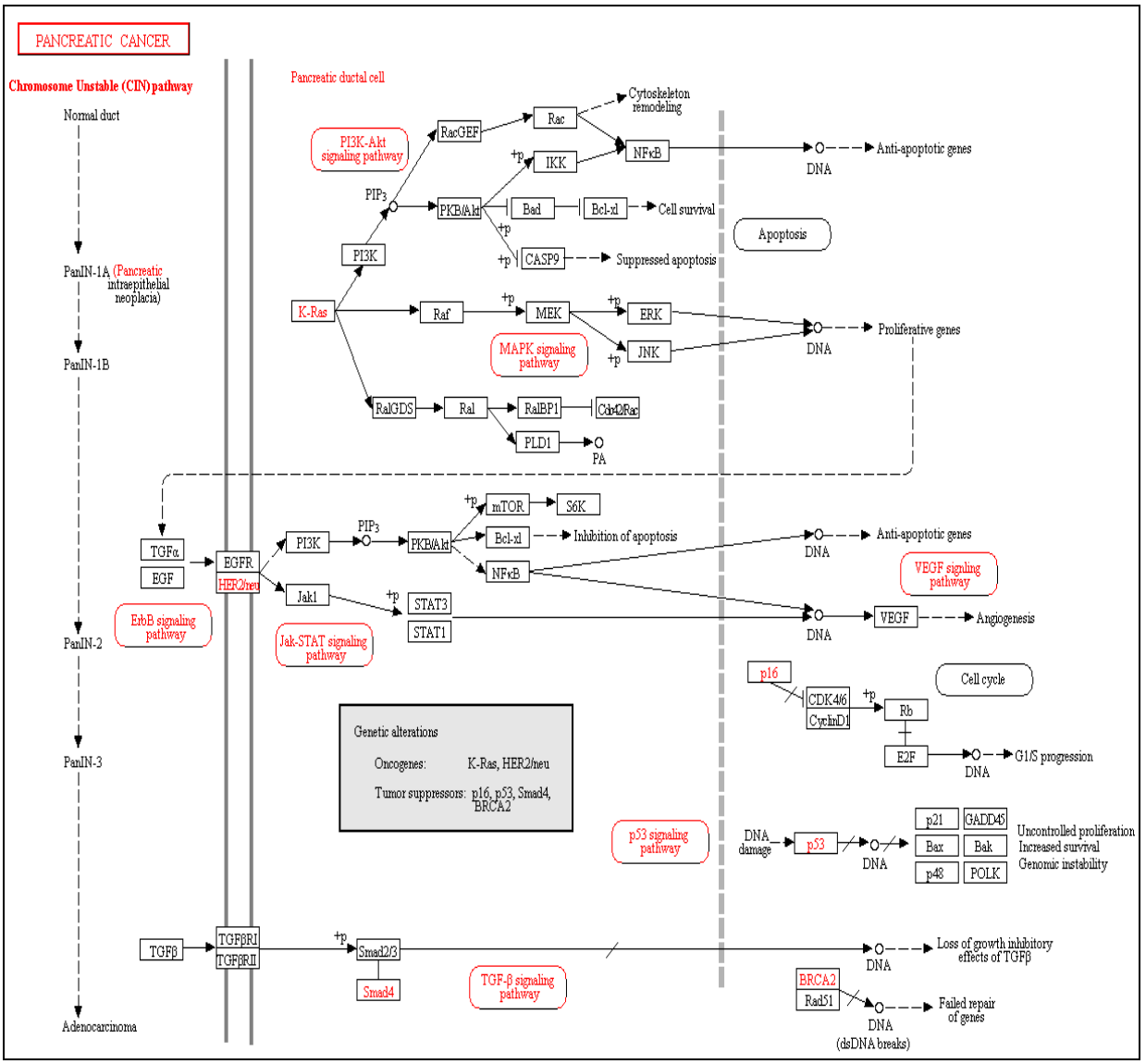


Fig24: KEGG Pathway for Pancreatic Cancer

We input all the quality IDs in DAVID gadget to consider viable clarifications and KEGG (Kyoto Encyclopedia of Qualities and Genomes) pathway investigation. Some pathways identified with growth are found in the perception in Table 14. Further, this pathway examination also reveals some more targets which are responsible for pancreatic infection, for example, central grip, GnRH flagging pathway, pancreatic growth pathway, renal cell carcinoma, chemokine hailing pathway

et cetera. We locate a couple of characteristics among KEGG pathways, for instance, FGS7, STAT5A, STAT5B, CDK2, FAS, MAP2K3, MMP2, MAPK3, KRAS, SOS2, ADCY4, ADCY7, GRB2, AGNT2, AKT3RALBP1, TP53, SMAD3, CCND1, RASSF5, PLCG2, PDPK1, which are in like manner in Pancreatic and other development cell lines and in Table 14. Among these characteristics a couple of characteristics like AKT3, CHUK, KRAS, TP53 are currently known as biomarkers for pancreatic development [46].

5.7 Biomarker Identification

We have completed our order on five PDAC quality articulation dataset using five existing classifiers and got ideal exactness in gathering classifier. We have furthermore analyzed our proposed technique on Pancreatic disease quality enunciation dataset and support the gained plans quantitatively. We in like manner perform utilitarian remark examination in DAVID instruments to get Gene Ontology remarks and furthermore KEGG pathways to recognize the Biomarkers for Pancreatic development. We found some tantamount characteristics which go about as Biomarkers [47]. Beside those characteristics we have discovered some more characteristics that are FAS, SPI1, GRB2, RASSF5, PLGC2 which may have basic part as Biomarkers for Pancreatic threat.

CHAPTER 6

❖Conclusion & Future Scope

6.1 Inference of the Thesis

6.2 Future Scope of the work

6 Conclusion & Future Scope

6.1 Inference of the Thesis

Recent DNA microarray technologies have made it possible to monitor transcription levels of tens of thousands of genes in parallel. Gene expression data generated by microarray experiments offer tremendous potential for advances in molecular biology and functional genomics.

Here, we reviewed various clustering algorithms, which have been applied to gene expression data onGSE15471seriesmatrix,GSE28735seriesmatrix,GSE32676seriesmatrix,GSE41368seriesmatrix and GSE71989seriesmatrix with promising results. Gene expression data can be clustered on both genes and samples.

Next, we classified the gene expression dataset through Ensemble method, where we achieved best result in Bagged tree classifier for Ensemble approach with 96.48% accuracy and in ANN classifiers, we got best result for Stateful GRU classifier with 99.22% accuracy.

Researchers typically select a few candidate algorithms and compare the clustering results. Nevertheless, we have shown that there are different aspects of cluster validation, and for each aspect, various approaches can be used to assess the quality or reliability of the clustering results. In fact, the performance of different clustering algorithms and different classification approaches is strongly dependent on both data distribution and application requirements. The choice of the clustering algorithm and classifying model is often guided by a combination of evaluation criteria and the user's experience.

After qualitative and statistical evaluation, we analyzed transcriptome profiles through functional enrichment relationship and KEGG pathway of optimized gene expression dataset using DAVID tool to identify biomarkers. Among them AKT3, TP53, CHUK, KRAS etc. have already discovered in previous work. Beside those characteristics we have discovered some more characteristics that FAS, SPI1, GRB2, RASSF5, PLGC2, CUL2, RASS5, GNA13, ADCY7, SMAD3 which we have seen repeatedly in our three different experiment. So, we can conclude here, these biomarkers also have greater influence in pancreatic cancer threat.

6.2 Future scope of the work

A gene expression data set typically contains thousands of genes. However, biologists often have different requirements on cluster granularity for different subsets of genes. For some purpose, biologists may be particularly interested in some specific subsets of genes and prefer small and tight clusters. While for other genes, people may only need a coarse overview of the data structure. Clustering is generally recognized as an unsupervised learning problem. Prior to undertaking a clustering task, global information regarding the data set, such as the total number of clusters and the complete data distribution in the object space, is usually unknown. However, some partial knowledge is often available regarding a gene expression data set.

For example, the functions of some genes have been studied in the literature, which can provide guidance to the clustering and classification. Furthermore, some groups of the experimental conditions are known to be strongly correlated, and the differences among the cluster structures under these different groups may be of particular interest. If a clustering algorithm could integrate such partial knowledge as some clustering constraints when carrying out the clustering task, we can expect the clustering results would be more biologically meaningful. In this way, clustering could cease to be a pure unsupervised process and become an interactive exploration of the data set. Similarly, we can expect the best result from classifiers.

We will find another way using deep learning approach with these series matrixes in future for a comparative study and to identify the genetic Biomarkers for PDAC.

Bibliography

[1] Hariharan, Deepak, A. Saied, and H. M. Kocher. "Analysis of mortality rates for pancreatic cancer across the world." *Hpb*10.1 (2008): 58-62.

[2] What you need to know about cancer of the pancreas. (14 July 2010). The National Cancer Institute (NCI). Booklet: NIH Publication No. 10-1560. http://www.cancer.gov/cancertopics/wyntk/pancreas.

[3] Hruban, Ralph H., et al. "Progression model for pancreatic cancer." *Clinical cancer research* 6.8 (2000): 2969-2972.

[4] Badea, Liviu, et al. "Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia-the authors reported a combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia." *Hepato-gastroenterology*55.88 (2008): 2016.

[5] Zhang, Geng, et al. "DPEP1 inhibits tumor cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic ductal adenocarcinoma." *PloS one* 7.2 (2012): e31507.

[6] Donahue, Timothy R., et al. "Integrative survival-based molecular profiling of human pancreatic cancer." *Clinical Cancer Research* 18.5 (2012): 1352-1363.

[7] Yin, Shuai, et al. "MiRNAs are Unlikely to be Involved in Retinoid Receptor Gene Regulation in Pancreatic Cancer Cells." *Cellular Physiology and Biochemistry* 44.2 (2017): 644-656.

[8] Olliffe I.T. *Principal Component Analysis*, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4.

[9] Dunn, Joseph C. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." (1973): 32-57.

[10] Herwig, Ralf, et al. "Large-scale clustering of cDNA-fingerprinting data." *Genome research* 9.11 (1999): 1093-1105.

[11] Meila, Marina, and Jianbo Shi. "Learning segmentation by random walks." *Advances in neural information processing systems*. 2001.

[12] Dönmez, Pınar. "Introduction to Machine Learning, by Ethem Alpaydın. Cambridge, MA: The MIT Press2010. ISBN: 978-0-262-01243-0. $54/£ 39.95+ 584 pages." *Natural Language Engineering* 19.2 (2013): 285-288.

[13] Olshausen, Bruno A., and David J. Field. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images." *Nature* 381.6583 (1996): 607.

[14] Mishra, Alok, and Mukesh Verma. "Cancer biomarkers: are we ready for the prime time?." *Cancers* 2.1 (2010): 190-208.

[15] Beuran, Mircea, et al. "The epithelial to mesenchymal transition in pancreatic cancer: a systematic review." *Pancreatology* 15.3 (2015): 217-225.

[16] Piva, Francesco, et al. "Epithelial to mesenchymal transition in renal cell carcinoma: implications for cancer therapy." *Molecular diagnosis & therapy* 20.2 (2016): 111-117.

[17] Takai, Erina, and Shinichi Yachida. "Genomic alterations in pancreatic cancer and their relevance to therapy." *World journal of gastrointestinal oncology* 7.10 (2015): 250.

[18] Majumder, Shounak, Suresh T. Chari, and David A. Ahlquist. "Molecular detection of pancreatic neoplasia: Current status and future promise." *World Journal of Gastroenterology: WJG*21.40 (2015): 11387.

[19] Andrikou, Kalliopi, et al. "Lgr5 expression, cancer stem cells and pancreatic cancer: results from biological and computational analyses." *Future Oncology* 11.7 (2015): 1037-1045.

[20] Harris, Midori A., et al. "The gene ontology project in 2008." *Nucleic Acids Research* 36.SUPPL. 1 (2008).

[21] García-Campos, Miguel A., Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. "Pathway analysis: state of the art." *Frontiers in physiology* 6 (2015): 383.

[22] Subramanian, Aravind, et al. "GSEA-P: a desktop application for Gene Set Enrichment Analysis." *Bioinformatics* 23.23 (2007): 3251-3253.

[23] Miller, Jeremy A., et al. "Strategies for aggregating gene expression data: the collapseRows R function." *BMC bioinformatics* 12.1 (2011): 322.

[24] Johnson, W. Evan, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics* 8.1 (2007): 118-127.

[25] Agrawal, Rakesh, et al. *Automatic subspace clustering of high dimensional data for data mining applications*. Vol. 27. No. 2. ACM, 1998.

[26] Dupuy, Alain, and Richard M. Simon. "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting." *Journal of the National Cancer Institute* 99.2 (2007): 147-157.

[27] Coombes, Kevin R., Jing Wang, and Keith A. Baggerly. "Microarrays: retracing steps." *Nature medicine* 13.11 (2007): 1276.

[28] Shi L, Campbell G, Jones WD, Campagne F, Wen Z, et al. (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 28: 827–838.

[29] Ntzani, Evangelia E., and John PA Ioannidis. "Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment." *The Lancet* 362.9394 (2003): 1439-1444.

[30] Yuan, Yuchen, et al. "DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations." *BMC bioinformatics* 17.17 (2016): 476.

[31] Meila, Marina, and Jianbo Shi. "Learning segmentation by random walks." *Advances in neural information processing systems*. 2001.

[32] Shi, Jianbo, and Jitendra Malik. "Normalized cuts and image segmentation." *IEEE Transactions on pattern analysis and machine intelligence* 22.8 (2000): 888-905.

[33] Meila, Marina, and Jianbo Shi. "Learning segmentation by random walks." *Advances in neural information processing systems*. 2001.

[34] Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm." *Advances in neural information processing systems*. 2002.

[35] Shannon, Claude E. "A mathematical theory of communications." *Bell Systems Technical Journal* 27 (1948): 379-423.

[36] Tan, Xiaodong, et al. "Phosphoproteome analysis of invasion and metastasis-related factors in pancreatic cancer cells." *PloS one* 11.3 (2016): e0152280.

[37] Liu, Peng, et al. "Quantitative secretomic analysis of pancreatic cancer cells in serum-containing conditioned medium." *Scientific Reports* 6 (2016): 37606.

[38] Opitz, David, and Richard Maclin. "Popular ensemble methods: An empirical study." *Journal of artificial intelligence research* 11 (1999): 169-198.

[39] Breiman, Leo. "Bagging predictors." *Machine learning* 24.2 (1996): 123-140.

[40] Aslam, Javed A., Raluca A. Popa, and Ronald L. Rivest. "On Estimating the Size and Confidence of a Statistical Audit." *EVT* 7 (2007): 8.

[41] Altman, Naomi S. "An introduction to kernel and nearest-neighbor nonparametric regression." *The American Statistician* 46.3 (1992): 175-185.

[42] Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, Brian P. (2007). "Section 16.5. Support Vector Machines". *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). New York: Cambridge University Press. ISBN 978-0-521-88068-8. Archived from the original on 2011-08-11.

[43] Lior, Rokach. *Data mining with decision trees: theory and applications*. Vol. 81. World scientific, 2014.

[44] Venables, W. N., and B. D. Ripley. "Random and mixed effects." *Modern applied statistics with S*. Springer, New York, NY, 2002. 271-300.

[45] Tan, Xiaodong, et al. "Phosphoproteome analysis of invasion and metastasis-related factors in pancreatic cancer cells." *PloS one* 11.3 (2016): e0152280.

[46] Liu, Peng, et al. "Quantitative secretomic analysis of pancreatic cancer cells in serum-containing conditioned medium." *Scientific Reports* 6 (2016): 37606.

[47] Pahari, Purbanka, Piyali Basak, and Anasua Sarkar. "Biomarker detection on Pancreatic cancer dataset using entropy based spectral clustering." *Research in Computational Intelligence and Communication Networks (ICRCICN), 2017 Third International Conference on*. IEEE, 2017.

[48] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555* (2014).

[49] Han, Jun, and Claudio Moraga. "The influence of the sigmoid function parameters on the speed of backpropagation learning." *International Workshop on Artificial Neural Networks*. Springer, Berlin, Heidelberg, 1995.

[50] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

[51] Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM." (1999): 850-855.

[52] Graves, Alex, et al. "A novel connectionist system for unconstrained handwriting recognition." *IEEE transactions on pattern analysis and machine intelligence* 31.5 (2009): 855-868.