A Survey on Machine Learning Approaches in Gene Expression Classification in Modelling Computational Diagnostic System for Complex Diseases

S. Karthik, M. Sudha

Abstract: ***In recent days, the survivability of people around the world has increased in a higher rate. The notable reason is the impact of the evolution of new technologies in medical systems that are invented to provide and improve healthcare for peoples. But still, there are some diseases, which may be identified also can be controlled. But there isn't any permanent solution for them such as cancer, psychiatric disorders etc. For those diseases, medical practitioners finds some way to discover medicine by analyzing the patient's genetic information such as DNA. Microarray technology is helpful in capturing biological genetic information to computer data. Computational techniques can be applied on those large set of genetic data of every individuals with or without disease, so that the genes that are responsible for the disease occurrence can be pointed out. Differentially Expressed Genes (DEG) are identified using many techniques. Machine Learning (ML) algorithms plays a significant role in identifying the distinction between normal genes and unhealthy genes, extracted from human genome. This paper is focusing on providing an in depth overview on different techniques on ML that are used to analyze and classifies the gene expression profiles of various diseases are discussed.***

Keywords: Gene Expression, Healthcare Systems, Machine Learning, Microarray data, Pattern Recognition.

I. INTRODUCTION

Microarray is a recent technology, helps to identify the patterns of gene expressions of multiple genes at a time in genomic level. It supports the researchers to analyze and investigate hundreds and thousands of genes in single experiment [32]. It detects various modern day diseases linked with every individual's genes like cancer, anemia etc. Gene Expression analysis provides a way to identify the genes that are expressed differentially [3], which are responsible for developing some diseases. Also, it shows the distinction between normal and abnormal genes using variety of mathematical models. Many publicly available dataset like Gene Expression Omnibus (GEO) [21], Array Express [11] etc. made the task easier to analyze gene patterns of many rare diseases. In recent days, there is an exponential development in medical field, globally.

Modern technology enhances the approach in developing advanced healthcare models like smart human health monitoring systems, personalized treatment etc. to diagnose everyone with at most care. Every day, everywhere, people are getting affected by various diseases,

Revised Manuscript Received on 28 December 2018.

S. Karthik, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore (Tamil Nadu), India.

Dr. M. Sudha, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore (Tamil Nadu), India.

Some may not be even diagnosed before it attains the critical level. Common diseases like cholera, malaria, dengue, common cold, fever can be easily diagnosed and cured with simple lab tests and available medications. But in case of cancer, psychiatric disorders and few diseases can be controlled in some stage with greater effort, sometimes cured only if it is diagnosed at initial stage itself.

Mostly, presence of these type of diseases are identified only after analyzing their biological samples such as tissues and cells. Mutation of genes is a typical reason for developing cancer [69] and it is cumbersome to invent a remedy to cure such diseases. Gene therapy [4] supports it by replacing the mutated genes with healthier one. But it will control the growth of abnormal cells rather destroying them completely. Modern computational systems helps the researchers to analyze complex data like genetic information of humans and its underlying patterns. These patterns reveals the genes that causes diseases. Mathematical models are helpful to build robust ML models for analyzing gene expression.

This paper is organized with the rest of the topics are given as sections as follows. Section 2 briefs the terms and concepts that are mostly used in this paper. Section 3 deals with the problem statement, background work of other researchers and current status of research in this field. Section 4 briefs some available microarray tools and databases.

In section 5, some related works on gene expression is discussed. Section 6 demonstrates gene expression analysis pathway that further extends with some steps that explains dimensionality reduction techniques, ML, its types and algorithms. Section 7 discusses few performance metrics. Section 8 reports some applications related to this field. Section 9 identifies the challenges faced in gene expression profiling. Section 10 discusses the future work and Section 11 concludes the review with summarization.

II. CONCEPTS AND TERMINOLOGIES

In this section, definitions of few terms that are highlighted in this paper are discussed briefly.

A. Biomarker

It is generally identified as the measurable biological indicator [14] that denotes the presence of the disease or its severity level. Some biomarkers are said as genes, molecules etc.

B. Differential Gene Expression

Some process, which determines the cells that are actively undergone transcription and translation to produce mRNA and proteins from genes under specific circumstances. DEG's helps to identify the distinction and variability between healthy and unhealthy cells.

C. Gene Expression Profiling

It is a method to figure out the genes in a cell that makes messenger RNA. These mRNA molecules carries the needed genetic information needed to produce different kind of proteins.

D. Microarrays

DNA Microarray technology is advanced and widely used in laboratories for genomic research. It helps the scientists to analyze between two different cells. It is performed by comparing every single gene between two cells on a single experiment. Microarrays are used for various purposes like gene interaction identification, toxicological drug research etc. [18].

III. BACKGROUND

A. Problem Statement

Gene Expression profiling is performed with various available statistical methods and modern ML techniques. Many researchers proposed different ways in analyzing microarray data. The values inside microarray data are probe intensities of expressed genes, which is in raw format. These values can be normalized using specialized normalization methods. After normalization, the probe data has genes as features and probe intensities of expressed genes as instances. Let, D be the identity of a dataset. Here, the total number of genes in dataset can be represented as G, where G = {g1, g2, g3…gn}. The total number of instances are said to be I, where I = {i1, i2, i3…in} and the classes are finite such as binary, ternary classification etc. The main part is identification of the DEG's, which are marked as informative genes. Many statistical approaches are profound which can very well suits for DEG analysis. After selecting the DEG's, the amount of features in D will be F, where F= {f1, f2, f3…fn}. Each feature has their own coordinates in D. (e.g.: First coordinate = {i1, g1} =f1). Biomarker identification is an important step before the process of classifying microarray data. These biomarkers can be selected as subset of F from the previously selected DEG's. This subset is considered as key features for next stage called classification. Robust ML algorithms could classify the data in best fit. Selection of ML algorithms needs clear analysis about the problem definition.

B. Related Work

Blood-based gene expression data of Autism Spectrum Disorder (ASD) is used to predict 42 samples taken from GEO [54]. They performed background correction with "normexp" function. Quantile method is applied for normalization. Supervised Learning algorithms such as Linear Discriminant Analysis (LDA), k-nearest neighbor (KNN) and Support Vector Machine (SVM) and is applied. Hierarchical clustering using Euclidean distance method and complete linkage is done for performing unsupervised learning. Among them, both SVM and KNN reached the accuracy level up to 93.8%, which is comparatively better than the results of LDA, which is 68.8%. Signal-to-noise ratio (SNR) technique is applied on the dataset to reduce its dimensions [48]. Out of 4026 genes, they selected only two significant genes for classification. Artificial Neural Network (ANN) algorithm classifies the data. This model achieved 93% accuracy.Gene Expression data of 53 patients with colon cancer were analyzed with different kind of neural network models [31]. They selected 500 genes out of all 54,675 genes from every individual persons. For classification process, S-Kohonen, SVM Neural Network and Back Propagation is used. Out of them, S-Kohonen outperformed other deployed classifiers by reaching 91% accuracy.

A new hybrid algorithm called Genetic Bee Colony (GBC) is proposed [1]. This algorithm combines Genetic Algorithm (GA) and Ant Bee Colony (ABC) optimization algorithm. This is used for selecting distinguished genes from all given genes. Three different datasets are tested. Along with this, three more multiclass classification datasets and all are binary classification SRBCT, lymphoma, and leukemia. SVM is used to classify the data. Colon dataset reached 98.39% and all other datasets achieved 100% classification accuracy from GBC-SVM model.To find the minimum set of genes from four disease dataset, Particle Swarm Optimization (PSO) and kNN is used [35]. This model reduces the computational time, selects best informative genes and gives better accuracy for blind test samples. Both MLL and SRBCT dataset reached 100% and ALL_AML reached 97.0588% accuracy for the given model.

A Fuzzy based model is proposed for classification of three gene expression disease datasets [34]. Entropy Filtering is applied to find most informative genes. This model produced better accuracy rates for all three datasets.A Feature Selection technique based on Correlation (CFS) is tested on schizophrenia dataset from GEO [77]. For mathematical modelling, many different ML algorithms are used, Locally Weighed Learning (LWL) algorithm delivered best result out from other learning models by reached 100% accuracy.

The presence of (CAPS2) gene is identified, which increases the risk factor for developing ASD [26]. The dataset is taken from GEO database repository. The experimentation is performed by utilizing various statistical methods and Geometric Particle Swarm Optimization (GPSO). GPSO – SVM model achieved 92.1% of accuracy from the experimental dataset.

Mutual Information (MI) technique is used in lymphoma and colon cancer and gene expression datasets to identify informative genes [70] for the prediction of bipolar disorder They used ANN, kNN and SVM with four different kernels namely linear, polynomial, quadratic and Radial Basis Function (RBF) for classification process. Leave One Out (LOOCV) Cross Validation is used for evaluating their model. This model reached the 95.98% accuracy for SVM polynomial kernel, which outperformed other classifiers.

T-Statistics, SNR and F-Statistics are used for ranking the genes [25]. Most informative genes are selected using Cuckoo Search optimization (CS). Five cancer datasets are used for experimentation. kNN is applied to dataset for classification, which is used as fitness function for CS. A higher accuracy of 100% is achieved by this model.

Partial Least Square based analysis has been made with Post-Traumatic Stress Disorder (PTSD) dataset collected from GEO repository. CytoScape tool to construct a network of DEG's [20]. Robust Multi Array technique is used for normalizing gene data. The result shows that, gene PRKCA has strong relation with PTSD. Also, other genes like CALM1, EP300 and TP53 are also has some partial co-relation with neurological disorders.

A two phase model is developed to classify liver cancer [57]. Initially, gene ranking has been performed using three different methods namely Enrichment score, Correlation and Analysis of Variance (ANOVA) to select most informative genes. SVM and Fuzzy Neural Network (FNN) classifiers are used in phase two. Enrichment score with FNN and Correlation with SVM combinations gives 100% accuracy separately.

Information Gain (IG) technique is used for selecting features and Standard Genetic Algorithm (SGA) is used to reduce the features [61]. For classification task, Genetic Programming technique is applied on dataset. This system is evaluated using seven different cancer datasets. K-fold cross validation is applied for validation of the model. Lung Michigan dataset reached 100% accuracy and other five datasets achieved accuracy around 90%.

An experimental work was conducted with GEO dataset (GSE 17612), which is mapped to schizophrenia disorder [47]. RMA is used for background correction. Quantile normalization is performed and median polish is done for summarization. DEG's are selected based on p value, if it is lesser than 0.01. The result from this study reveals that CCL3, S100A8, SYK and VEGFA are the genes mainly overlaps on schizophrenic patients as well as related with other psychological disorders. SVM with Recursive Feature Elimination (RFE) is applied for selecting best genes. A total of 21 genes are selected as best features. Random Forest (RF), Extremely randomized tree forests classifiers are developed to evaluate their performance, Out of them, RF has shown better result by achieving Area Under Curve (AUC) value 0.98.

A set of most DEG's are selected based on its p value ($p<0.01$) from schizophrenic disorder dataset [76]. Top 500 DEG's are selected as a subset of features from all the features in the dataset. Multilayer Perceptron (MLP) Neural Network classifier is applied on dataset to develop the model. An Analytical Framework is developed for the prediction of bipolar disorder [43]. A Filtering method called Significance Analysis of Microarrays (SAM) is used. RFE-SVM classifier is used. *NOG* and *CTBP1* are concluded as the genes, which are mainly responsible for bipolar disorder.

A genetic score based risk identification model for bipolar disorder is developed from genome wide association data [13]. They applied Random Forest algorithm to classify genes. From the classification rate, risk score is calculated. Weighed Multiple Logistic Regression (WMLR) classifier is developed to discriminate the patients with MDD [19]. For noise removal, Multiple Imputation technique is applied. No potential biomarkers are identified in this system.

C. Current Status of Gene Expression Analysis in Research

- Gene Expression data analysis to enable personalized medicine for complex genetic diseases [71].
- Identification of chronic drug effects on patient through gene expression modelling [74].
- Pathway analysis process and gene set enrichment [52]
- Gene biomarker identification for multiple complex disease [65].

IV. MICROARRAY TOOLS AND DATABASES

A curated set of databases are provided to perform scientific research and statistical computation from the data. Mostly the databases are structured and open source. Few are listed out below in Table I and Table II, not in any specific order.

Table I. List of Microarray Gene Expression Analysis Tools

Tool Name	License Type	Features	Reference Link
Bioconductor [23]	Open Source	Analysis and comprehension of high-throughput genomic data.	https://www.bioconductor.org/
Gene Pattern [59]	Open Source	Genomic Data Analysis	http://software.broadinstitute.org/cancer/software/genepattern/
CytoScape [64]	Open Source	Complex Network Visualization process and Data Integration	http://www.cytoscape.org/
BRB Array Tools. [63]	Free Excel Plug-in	Statistical analysis and data visualization.	https://brb.nci.nih.gov/BRB-ArrayTools/
Thermo Fisher Scientific	Commercial	Microarray Analysis	https://www.thermofisher.com/in/en/home/life-science/microarray-analysis.html

Table II. List of Freely Accessible Gene Expression Databases.

Database Name	Database Information	No. of sample profiles	Reference Link
Gene Expression Omnibus - NCBI	GEO is a curated, publicly available database supports MIAME compliances.	641770 (2011)	https://www.ncbi.nlm.nih.gov/geo/

ArrayExpress at EBI	ArrayExpress archives the data from genomic experiments.	708914 (2011)	https://www.ebi.ac.uk/arrayexpress/
The Cancer Genome Atlas (TCGA) [75]	TCGA database contains gene expression data for various types of cancer.	-	https://cancergenome.nih.gov/
ImmGen database	Immune gene data and co-regulation of mouse is stored for research purpose.	1059 (2012)	https://www.immgen.org/
GeneNetwork system	Exon arrays, RNA-Seq data especially for genetic analysis and storage	~10000 (2010)	http://www.genenetwork.org/webqtl/main.py

V. ANALYSIS OF EXISTING RELATED WORKS

Many Researchers contributed their works in biomarker identification using different methodologies. A collection of such similar articles are analysed and discussed with their methodologies in Table III given below.

Table III. Some Previously Published Works and its Results.

Authors	Dataset	Algorithms and Techniques	Accuracy
[62]	GSE12654	Naïve Bayes(NB)	87.71%
[2]	GSE9222	Conditional Mutual Information Maximization and SVM-RFE	89.50%
[68]	Colon, leukemia and breast cancer	Ensemble Model	0.99, 1 and 1 (AUC)
[50]	Colon, ALL_AML, SRBCT, MLL, Tumors_9 and Tumors_11	Genetic Algorithm and Learning Automata and SVM	99.46% 100%, 97.35% , 93.96%, 86.52%, and 84.38%.
[67]	Leukemia, colon and prostate	Iterative Transductive Support Vector Machine	97.8%, 86.2% and 93.6%
[24]	Breast cancer and leukemia	SVM	93.75% and 98%
[7]	Leukemia	RFE and Based Bayes error Filter (BBF) and SVM	95.833%
[28]	MLL Leukemia Colon DLBCL Prostate	ReliefF, Decision tree, k-nearest neighbor, SVM and random forest	99.89%, 99.40%,99.93 %, 99.69% and 99.35
[80]	Colon cancer dataset.	Bhattacharyya distance and SVM	90.5% and 96.8%
[6]	Leukemia	AdaBoost incorporating linear SVM (AdaSVM)	95.34%
[8]	Lymphoma data set and SRBCT	extreme Learning Machine (ELM)	89.89%, 85.67%
[10]	Leukemia, prostate cancer, colon cancer	Signal to Noise Ratio (SNR), k-NN, SVM	100%
[55]	Breast_b, CNS, Leukemia, Lymphoma, MLL_Leukemia and Prostate cancer	Ensemble Correlation-Based Gene Selection algorithm and SVM	100%
[66]	Leukemia and breast	RFE, SVM, Adaboost	96.22%
[36]	laryngeal, bladder and colorectal	Discriminative deep belief networks	93.3%
[22]	Leukemia, colon, lymphoma and prostate cancer	Cuckoo Optimization Algorithm (COA) and Genetic Algorithm (GA), SVM and MLP	98.9%, 99.2%, 88.9% and 91.5%
[16]	Leukemia, Prostate and SRBCT	Multi-objective binary bat algorithm with specific local searches	-
[45]	FHCCancer 9	Hierarchical Clustering, k-NN, SVM and RF	98.41%
[41]	Leukemia	SMO (Sequential Minimal Optimization) Algorithm and SVM	94.11%
[58]	colon, leukemia, and prostate cancers	random forest (RF)	87.38%, 73.33% and 95.23%
[46]	Leukemia, Lymphoma and SRBCT	Hybrid Back Propagation Neuro fuzzy Method (BPN) and (RKLM)	98%, 96%, 97%

[60]	Colon Leukemia Lymphoma GCM	naïve Bayes, instance-based and decision trees, BIRS	85.48%, 93.04%, 67.37%
[79]	ALLAML3 C, DLBCL_A, SRBCT,	Weight Local Modularity (*WLM*), k-nearest neighborhood classifier and SVM	100%
[12]	colon cancer	Monte Carlo algorithm, PCA	84%
[5]	Breast cancer	Hybrid Ensemble Gene Selection () Algorithm	96.9%

VI. GENE EXPRESSION ANALYSIS USING DIFFERENT MACHINE LEARNING APPROACHES

In the pathway of analyzing any gene expression data, it is inevitable to undergo some sort of processing. Those steps are almost necessary to develop better models like pre-processing of data (cleaning, wrangling), dimensionality reduction, hyper-parameter tuning, gene selection, cross-validation and classification or clustering of data. Out of these, two important steps namely dimensionality reduction methods and few classification algorithms, which are described in later sections. A typical architecture diagram of microarray data processing and classification is described in Fig. I.

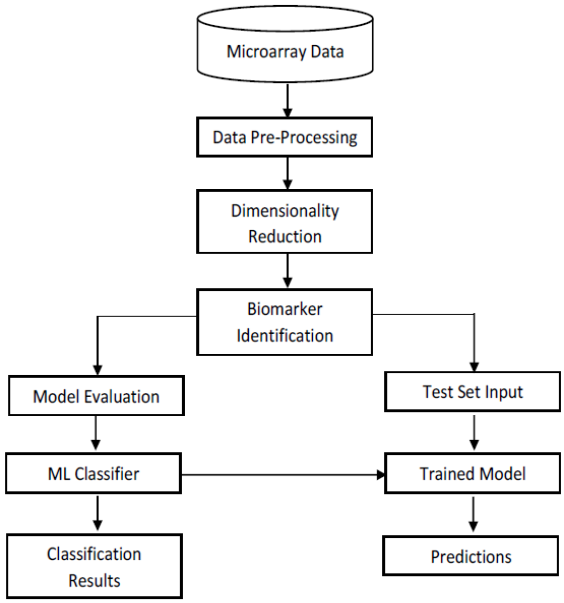

Fig. 1. Microarray Data Classification Process using Machine Learning

An important challenge encountered during the processing of gene expression is its "dimension" (i.e. size of data) [30]. It is not a good practice to use all the gene features given in the dataset. Only the informative gene subset can be selected to train a model. To do that, some techniques are available, which performs better with gene expressions. Mainly dimensionality reduction is classified into two sub divisions namely feature selection and feature extraction.

A. Feature Selection

It is an important process in ML model determines its robustness. It mainly reduces the over fitting of data. Also, it improves model accuracy [37].

a. Filter Method

In this method, the features are selected with respect to its "relevance" [73]. Mostly, univariate statistical analysis are performed for the evaluation of the features. It identifies correlation between features and the outcome variable. Some examples are fisher, correlation coefficient, chi-square, information gain etc.

b. Wrapper Method

Initially, a subset of features are trained using any ML model. Based on its outcome, some features are retained for next iteration and other features are eliminated. Based on cross-validation the features are evaluated. This method improves performance by selecting useful features. But it is computationally expensive, since it trains the subset of data for multiple iterations [9]. Example for this method are Forward Selection, RFE, and Backward Elimination etc.

c. Embedded Method

It integrates the properties of both wrapper method and filter method. This method uses the algorithms that has some pre-defined methods of feature selection by itself. Examples for this method are LASSO and RIDGE regression models that uses regularization and penalizing technique to reduce over-fitting.

d. Feature Extraction

The transformation process of input data to some set of useful features called as feature vectors are selected based on the relevancy to the system. It would be much helpful to represent a better learning model. Some notable properties of evaluating feature extraction methods are, it should look for the feature which are non-redundant, informative and more generalizing. There are some advantages from this methods. Multi-collinearity problem will be solved [53]. Also it improves system performance, reduces noise and simplifies the plotting of data in dimensional space.

e. Principal Component Analysis (PCA)

PCA is a linear transformation technique uses orthogonal transformation technique to transform the input features with possible correlation into uncorrelated features [42]. The transformed features are called as principle components. It projects the data from a high dimensional space into low dimension space. It finds the axes with most variance and data that spreads over the space.

f. Kernel – PCA

This method uses "kernel-trick" to handle non-linearly separable data. It finds the nonlinear pattern from the data [78].

g. Auto-Encoders

To produce the representation of higher dimensional data into lower dimension, an unsupervised learning technique based on neural network called "auto-encoders" is used. The limitations of linear PCA is rectified in this method by supports non-linearity in data. It has two units called encoder and decoder. Encoder is used in training and evaluation process were decoder is only in training phase.

It identifies the compressed view of information from the given data. Decoder mirrors the encoder network whose purpose is to reconstruct the network based onoriginal data. At last, the informative features are sorted out in compressed representation and it shows much better results than PCA [29].

VII. MACHINE LEARNING TECHNIQUES

Machine Learning is a sub-field of Artificial Intelligence [39]. It gives the ability to amachine to learn by experience, without being programmed explicitly with the help ofvarious mathematical models, logical functions and huge amount of real time data. It can beapplied in many business areas such as finance, marketing, sales, government andhealthcare sectors. Many machine learning algorithms are developed for handling differentkind of problems. Predictive Analytics applies machine learning algorithm to analyzepurchasing pattern, loan assessment, fraud detection etc. [38]. Also, recommendation engines,sentiment analysis social media trend analysis systems are developed with the help ofvarious machine learning models. The role of machine learning in current scenario is veryhigh due to its resilience among various sectors and the impact created after involvement ofthese models in solving real world problems gives a new dimension of technologicaldevelopment. In future, it plays a vital role in most part of every individual's activities over internet. Also, it is considered as the most important factor in the advancements oncomputer vision and intelligent computing of this century.

A. Supervised Learning

The process of mapping the training set of input and output samples where all thedata are labeled [40]. Some of the widely used classification algorithms are Decision Trees,Support Vector Machine, k-Nearest Neighbor, Artificial Neural Network, Naive Bayesalgorithm, Logistic Regression etc.

B. Unsupervised Learning

Unsupervised learning gives, how well the systems are able to learn to represent any particular input patterns in a way it reflects the statistical pattern of the overall input. Unlike supervised learning, target variable will not be present in this learning method[17].

C. Semi-Supervised Learning

It performs the learning task from few labelled and huge unlabeled data [81]. So, informally it is said that it may do both supervised/unsupervised way of learning. The main intention of this method is to classify the data that is unlabeled using the labelled set.

D. Reinforcement Learning

An agent or algorithm, which can learn by itself based on its interaction with the environment. Also it performs some actions, receives punishments and rewards depends on response it perceives from the situation [44]. The intention of this system is to maximize its performance by learning in new environment.

E. Machine Learning Algorithms

a. Linear Regression

Linear Regression observes and analyses the past information about the relationship between variables and data [49]. From the observation, it would predict what may occur in future. There should be a independent and dependent variable needed for applying regression. If two or more independent variables used, then it is called as multiple regression, which may act as both linear and non-linear. Independent variable is called as exploratory variable.

b. Support Vector Machines

SVM is a supervised learning algorithm [72]. It divides the data using an optimal hyper-plane. Here, kernel plays a major role, is a similarity function. It takes the input, separate it with similarity measure between each of them. Linear, Gaussian, RBF and Polynomial are some types of kernels used frequently. Kernel selection reflects in model accuracy. It solves both single and multi-class classification type of data. It will achieve better results with small datasets, but it depends on tuning the parameters and choosing the correct kernel for the data.

c. Decision Tree

A decision tree is a tree based model [56], generates all possible chances of occurrence of events and its consequences by observing logical connection between each features in dataset. This algorithm itself splits into different branches using some better techniques to generate subsets of logical sets from the given sample. It classifies the input well with fewer amounts of samples. This algorithm generalizes the system in good way that helps understanding the logical working of system directly. Some notable DT algorithms are ID3, CHAID and C4.5.

d. Random Forest

Random Forest is an ensemble learning method. It follows the divide and conquer method to improve overall performance. Also, the model aims to boost up the weak learners into strong learner. Every single learner is assumed as weak learner, but together will form strong learners [15]. This is used on classification and regression type of problems. It will construct a group of decision trees to train data with individual trees. The result will be obtained by calculating the mean value returned by all the decision trees. This algorithm works well with the dataset having more number of predictive variables and samples due to high resilience of observing the variance of every tree and uses huge amount of samples to be involved in training the model.

e. Artificial Neural Network

ANN's are inspired by biological neural networks, acts as a statistical learning model, mainly used in ML for developing predictive models. These networks are constructed by interconnecting a large number of neurons that communicate between each other. A typical neural network contains input layer, output layer and hidden layer constructed with nodes and some components such as bias, weights, learning rate,

Activation function and so on [33]. It uses forward propagation to compute the output for the first pass where weights are assigned randomly. In back propagation, error rate margin will be calculated and weights are adjusted to lower the error rate in the next pass. The same will be repeated till the system reaches optimal result or the sufficient epochs completed. One complete cycle of traversing through entire dataset is called as an epoch.

It is used to solve more real world problems. Its complex structure by nature, versatility and resilience all together gives some way to create "Deep Learning" Algorithms. These algorithms are more reliable for handling complex systems, especially for images. It provides flexibility to handle any kind of data.

f. K-Means Clustering

It comes under unsupervised ML technique. It is used when the dataset is unlabeled. It group the data into multiple clusters, which is represented by variable 'k'. The data points assigned to any one of the clusters iteratively that depends on the similarity between the features. The main part is, the number of cluster to be generated is initialized before applying to the data [27].

VIII. PERFORMANCE EVALUATION METRICS

Measuring performance of ML model lies under few validation metrics. Confusion matrix is an important measure mainly used in classification models that calculates four different factors called true positive (tp), false positive (fp), true negative (tn) and false negative (fn). It finds the number of correctly and incorrectly classified instances from the samples given to test the model [37]. Performance metrics and its formula are discussed in Table IV given below. Few performance metrics are discussed below in detail.

Accuracy of a model is calculated using four measures called fp, fn, tp and tn.

tp – The result of tp identifies the condition when it is present.
fp – The result of fp identifies the condition when it is not present.
tn – The result of fp doesn't find the condition when it is not present.
fn – The result of fn doesn't find the condition when it is present.

Sensitivity calculates the proportion of correctly identified instances with actual positives. Specificity finds the proportion of correctly identified instances with actual negatives. To find the test accuracy of the model, F-score is calculated [37]. Root Mean Squared Error (RMSE) is the mean of squared difference between observation and prediction with a square root. The average magnitude of error can be calculated using RMSE.

Table IV. Performance Metrics

Performance Metrics	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
f-Score	$\frac{(\beta^2+1)*precision*recall)}{\beta^2*precision+recall}$
Root Mean Squared Error	$\sqrt{\sum_{i=1}^{n}\frac{(Pi-Oi)^2}{n}}$

IX. APPLICATIONS

Gene Expression Analysis provides a better path to identify DEG's. The urge to finding those genes are helpful to use it in the form of developing various applications like personalized treatment, disease diagnosis, gene discovery, drug discovery (drug target identification, drug target interaction) [51], tumor classification etc. ML helps in finding the patterns and the distinction between the data. It owns great algorithms as tools that is applied on various fields.

X. FACTORS CONTRIBUTING TO THE EXISTING PROBLEM METHODOLOGIES

There are lot of advantages in analyzing gene expressions that enhances quality of providing better healthcare to people. But still some issues are found in the system that are to be highlighted.

- Only a handful number of researchers are actively working on creating curated and structural databases for various complex diseases. So, chances of creating powerful computational models will lags due to the lack of data samples.
- Genetic data is very complex and large in size. So, to handle those large data, advanced storage systems are needed.
- Simple math models will not work well for identifying patterns from gene data. Also, to analyse and observe the data, advance processing devices are needed. They should be capable of performing modelling on the data.
- Gap between researchers, biologist and medical practitioner's reduces the chance of discovering new methods and techniques for finding solutions to any domain-specific problems.

XI. FUTURE WORK

- In future, there will be humongous amount of genetic data to be gathered from large sample of people for various diseases, which may be useful in identifying the correlation between malignant and normal genes for every specific disease.
- Also, high performance computational systems will be dedicated to analyse every individual's genetic information so that

It would be possible to predict chances of occurrence of diseases in future based on family history or environmental factors.

- Personalized Treatment for every individual is also possible by sorting out their gene patterns from other data. Right medicine for the malignant gene will reduce risk of developing diseases.

XII. CONCLUSION

In this paper, a comprehensive review on gene expression analysis using various ML techniques are discussed. The outcome of this review reveals the importance of ML in gene expression profiling for finding novel biomarkers for various complex diseases. Following that, advanced research activities on microarray data processing are highlighted. Varieties of learning algorithms and its evaluation metrics are underlined. Analyzing gene expression profiles reveals valuable insights that would be helpful to identify DEG's and abnormal genes, responsible for the developing various diseases. Effective ways to find the cure for genetic disorders with advance technologies and gene sequencing techniques namely microarrays, next generation sequencing, ML, pattern recognition and other high through-put computational methods paves a way to develop reliable computational models to diagnose complex diseases that will hopefully increase the chance of living without disease.

Summary - This paper represents the need for gene expression profiling and its methods using various ML techniques. A lot of previous research work done by other researchers in this field are discussed briefly. But still, there is some gap identified in this area. Few are said to be pharmacogenomics studies, precision medicine discovery and psychiatric disorder pathway analysis through gene expressions. In future, ML models will opens up many opportunities to discover new insights from the about discussed research gaps. In specific, analysis of gene expressions of genetic disorders such as any psychiatric disorders, cancer and other deadly diseases will pave a new way to discover new medicines and therapies to cure them permanently. In addition to ML algorithms, other techniques like nature inspired algorithms, swarm intelligent algorithms will also be much helpful in developing robust gene expression analysis models.

ACKNOWLEDGEMENTS

Funding – No funding for this research work

Compliance with Ethical Standards:

This article does not contain any studies with human participants performed by any of the authors.

REFERENCES

1. Alshamlan, H.M., Badr, G.H. and Alohali, Y.A., 2015. Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Computational biology and chemistry, 56, pp.49-60.
2. Alzubi, R., Ramzan, N. and Alzoubi, H., 2017, August. Hybrid feature selection method for autism spectrum disorder SNPs. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2017 IEEE Conference on(pp. 1-7). IEEE.
3. Ansel, A., Rosenzweig, J.P., Zisman, P.D., Melamed, M. and Gesundheit, B., 2017. Variation in gene expression in autism spectrum disorders: an extensive review of transcriptomic studies. Frontiers in neuroscience, 10, p.601.
4. Aragona, M. and Blanpain, C., 2017. Gene therapy: Transgenic stem cells replace skin. Nature, 551(7680), p.306.
5. Aruna, d. And nandakishore, d., 2013. Hybrid ensemble gene selection algorithm for identifying biomarkers from breast cancer gene expression profiles.
6. Begum, S., Chakraborty, D. and Sarkar, R., 2015, December. Cancer classification from gene expression based microarray data using SVM ensemble. In Condition Assessment Techniques in Electrical Systems (CATCON), 2015 International Conference on (pp. 13-16). IEEE.
7. Bennet, J., Ganaprakasam, C. and Kumar, N., 2015. A Hybrid Approach for Gene Selection and Classification using Support Vector Machine. International Arab Journal of Information Technology (IAJIT), 12.
8. Bharathi, A. and Natarajan, A.M., 2010, December. Microarray gene expression cancer diagnosis using Machine Learning algorithms. In Signal and Image Processing (ICSIP), 2010 International Conference on (pp. 275-280). IEEE.
9. Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J.M. and Herrera, F., 2014. A review of microarray datasets and applied feature selection methods. Information Sciences, 282, pp.111-135.
10. Bouazza, S.H., Hamdi, N., Zeroual, A. and Auhmani, K., 2015, March. Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers. In Intelligent Systems and Computer Vision (ISCV), 2015 (pp. 1-6). IEEE.
11. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G. and Oezcimen, A., 2003. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic acids research, 31(1), pp.68-71.
12. Chen, H., Zhao, H., Shen, J., Zhou, R. and Zhou, Q., 2015, June. Supervised machine learning model for high dimensional gene data in colon cancer detection. In Big Data (BigData Congress), 2015 IEEE International Congress on(pp. 134-141). IEEE.
13. Chuang, L.C. and Kuo, P.H., 2017. Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. Scientific Reports, 7, p.39943.
14. Cross, A.J., Lampe, J.W., Rock, C.L. and Boushey, C.J., 2017. Biomarkers and their use in nutrition intervention. In Nutrition in the Prevention and Treatment of Disease (Fourth Edition)(pp. 217-234).
15. Cutler, A., Cutler, D.R. and Stevens, J.R., 2012. Random forests. In Ensemble machine learning (pp. 157-175). Springer, Boston, MA.
16. Dashtban, M., Balafar, M. and Suravajhala, P., 2018. Gene selection for tumor classification using a novel bio-inspired multi-objective approach. Genomics, 110(1), pp.10-17.
17. Dayan, P., Sahani, M. and Deback, G., 1999. Unsupervised learning. *The MIT encyclopedia of the cognitive sciences*.
18. de Freitas, R.C.C., Bortolin, R.H., Lopes, M.B., Hirata, M.H., Hirata, R.D.C., Silbiger, V.N. and Luchessi, A.D., 2016. Integrated analysis of miRNA and mRNA gene expression microarrays: Influence on platelet reactivity, clopidogrel response and drug-induced toxicity. Gene, 593(1), pp.172-178.
19. Dipnall, J.F., Pasco, J.A., Berk, M., Williams, L.J., Dodd, S., Jacka, F.N. and Meyer, D., 2016. Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. PloS one, 11(2), p.e0148195.
20. Dong, K., Zhang, F., Zhu, W., Wang, Z. and Wang, G., 2014. Partial least squares based gene expression analysis in posttraumatic stress disorder. Eur Rev Med Pharmacol Sci, 18(16), pp.2306-2310.
21. Edgar, R., Domrachev, M. and Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), pp.207-210.
22. Elyasigomari, V., Mirjafari, M.S., Screen, H.R. and Shaheed, M.H., 2015. Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization. Applied Soft Computing, 35, pp.43-51.
23. Gentleman, R., Carey, V., Huber, W., Irizarry, R. and Dudoit, S. eds., 2006. Bioinformatics and computational biology solutions using R and Bioconductor. Springer Science & Business Media.
24. Gour, D.K., Jain, Y.K. and Pandey, G.S., 2011. The Classification of Cancer Gene using Hybrid Method of Machine Learning. International Journal of Advanced Research in Computer Science, 2(2).

25. Gunavathi, C. and Premalatha, K., 2015. Cuckoo search optimisation for feature selection in cancer classification: a new approach. International journal of data mining and bioinformatics, 13(3), pp.248-265.
26. Hameed, S.S., Hassan, R. and Muhammad, F.F., 2017. Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a GBPSO-SVM algorithm. PloS one, 12(11), p.e0187371.
27. Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), pp.100-108.
28. Hijazi, H. and Chan, C., 2013. A classification framework applied to cancer gene expression profiles. Journal of healthcare engineering, 4(2), pp.255-283.
29. Hinton, G.E. and Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. science, 313(5786), pp.504-507.
30. Hira, Z.M. and Gillies, D.F., 2015. A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, 2015.
31. Hu, H.P., Niu, Z.J., Bai, Y.P. and Tan, X.H., 2015. Cancer classification based on gene expression using neural networks. Genetics and Molecular Research, 14(4), pp.17605-17611.
32. Huynh, P.H., Nguyen, V.H. and Do, T.N., 2018. A Coupling Support Vector Machines with the Feature Learning of Deep Convolutional Neural Networks for Classifying Microarray Gene Expression Data. In Modern Approaches for Intelligent Information and Database Systems (pp. 233-243). Springer, Cham.
33. Jain, A.K., Mao, J. and Mohiuddin, K.M., 1996. Artificial neural networks: A tutorial. Computer, 29(3), pp.31-44.
34. Kalaiselvi, N. and Inbarani, H.H., 2013. Fuzzy soft set based classification for gene expression data. arXiv preprint arXiv:1301.1502.
35. Kar, S., Sharma, K.D. and Maitra, M., 2015. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Systems with Applications, 42(1), pp.612-627.
36. Karabulut, E.M. and Ibrikci, T., 2017. Discriminative deep belief networks for microarray based cancer classification. Biomedical Research, 28(3).
37. Karthik, S., Perumal, R.S. and Mouli, P.C., 2018. Breast Cancer Classification Using Deep Neural Networks. In Knowledge Computing and Its Applications (pp. 227-241). Springer, Singapore.
38. Kelleher, J.D., Mac Namee, B. and D'Arcy, A., 2015. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press.
39. King, B.F., 2018. Artificial Intelligence and Radiology: What Will the Future Hold?. Journal of the American College of Radiology.
40. Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, pp.3-24.
41. KR, S., 2011. Microarray data classification using support vector machine. International Journal of Biometrics and Bioinformatics (IJBB), 5(1), p.10.
42. Kumar, G. and Bhatia, P.K., 2014, February. A detailed review of feature extraction in image processing systems. In Advanced Computing & Communication Technologies (ACCT), 2014 Fourth International Conference on (pp. 5-12). IEEE.
43. Leska, V., Bei, E.S., Petrakis, E. and Zervakis, M., 2016. Gene Expression Data Analysis for Classification of Bipolar Disorders. In XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016 (pp. 500-506). Springer, Cham.
44. Littman, M.L., 2015. Reinforcement learning improves behaviour from evaluative feedback. Nature, 521(7553), p.445.
45. Liu, Y.X., Zhang, N.N., He, Y. and Lun, L.J., 2015. Prediction of core cancer genes using a hybrid of feature selection and machine learning methods. Genetics and Molecular Research, 14(3), pp.8871-8882.
46. Loganathan, C. and Girija, K.V., Investigations on Hybrid Learning in Anfis in Microarray Gene Expression Classification.
47. Logotheti, M., Pilalis, E., Venizelos, N., Kolisis, F. and Chatziioannou, A., 2016. Studying Microarray Gene Expression Data of Schizophrenic Patients for Derivation of a Diagnostic Signature through the Aid of Machine Learning. Biom Biostat Int J, 4(5), p.00106.
48. Mehridehnavi A, Ziaei L. Minimal gene selection for classification and diagnosis prediction based on gene expression profile. Advanced Biomedical Research. 2013;2:26. doi:10.4103/2277-9175.107999.
49. Montgomery, D.C., Peck, E.A. and Vining, G.G., 2012. Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.
50. Motieghader, H., Najafi, A., Sadeghi, B. and Masoudi-Nejad, A., 2017. A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. Informatics in Medicine Unlocked, 9, pp.246-254.
51. Nath, A., Kumari, P. and Chaube, R., 2018. Prediction of Human Drug Targets and Their Interactions Using Machine Learning Methods: Current and Future Perspectives. In Computational Drug Discovery and Design (pp. 21-30). Humana Press, New York, NY.
52. Nie, Y., Chen, V., Shannon, C.P., Andiappan, A.K., Lee, B., Rotzschke, O., Castaldi, P.J., Hersh, C.P., Fishbane, N., Ng, R.T. and McManus, B., 2017. Network-based analysis reveals novel gene signatures in peripheral blood of patients with chronic obstructive pulmonary disease. *Respiratory research*, *18*(1), p.72.
53. Nilashi, M., bin Ibrahim, O., Ithnin, N. and Sarmin, N.H., 2015. A multi-criteria collaborative filtering recommender system for the tourism domain using Expectation Maximization (EM) and PCA–ANFIS. Electronic Commerce Research and Applications, 14(6), pp.542-562.
54. Oh, D.H., Kim, I.B., Kim, S.H. and Ahn, D.H., 2017. Predicting Autism Spectrum Disorder Using Blood-based Gene Expression Signatures and Machine Learning. Clinical Psychopharmacology and Neuroscience, 15(1), p.47.
55. Piao, Y., Piao, M., Park, K. and Ryu, K.H., 2012. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. Bioinformatics, 28(24), pp.3306-3315.
56. Quinlan, J.R., 1986. Induction of decision trees. Machine learning, 1(1), pp.81-106.
57. Rajeswari, P. and Reena, G.S., 2011. Human liver cancer classification using microarray gene expression data. International Journal of Computer Applications, 34(6), pp.25-37.
58. Ram, M., Najafi, A. and Shakeri, M.T., 2017. Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest. Iranian Journal of Pathology, 12(4), pp.339-347.
59. Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P. and Mesirov, J.P., 2006. GenePattern 2.0. Nature genetics, 38(5), p.500.
60. Ruiz, R., Riquelme, J.C. and Aguilar-Ruiz, J.S., 2006. Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognition, 39(12), pp.2383-2392.
61. Salem, H., Attiya, G. and El-Fishawy, N., 2017. Classification of human cancer diseases by gene expression profiles. Applied Soft Computing, 50, pp.124-134.
62. Saylan, C.C. and Yılancıoğlu, K., 2016. Classification of Schizophrenia and Bipolar Disorder by Using Machine Learning Algorithms. The Journal of Neurobehavioral Sciences, 3(3), pp.92-95.
63. Simon, R., Lam, A., Li, M.C., Ngan, M., Menenzes, S. and Zhao, Y., 2007. Analysis of gene expression data using BRB-array tools. Cancer informatics, 3, p.117693510700300022.
64. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. and Ideker, T., 2010. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics, 27(3), pp.431-432.
65. Soe, H.J., Yong, Y.K., Al-Obaidi, M.M.J., Raju, C.S., Gudimella, R., Manikam, R. and Sekaran, S.D., 2018. Identifying protein biomarkers in predicting disease severity of dengue virus infection using immune-related protein microarray. *Medicine*, *97*(5).
66. Song, N., Wang, K., Xu, M., Xie, X., Chen, G. and Wang, Y., 2015. Design and analysis of ensemble classifier for gene expression data of cancer. Journal of Clinical & Medical Genomics, pp.1-7.
67. Tajari, H. and Beigy, H., 2012. Gene Expression Based Classification using Iterative Transductive Support Vector Machine. International Journal of Machine Learning and Computing, 2(1), p.76.
68. Tarek, S., Elwahab, R.A. and Shoman, M., 2017. Gene expression based cancer classification. Egyptian Informatics Journal, 18(3), pp.151-159.
69. Tomasetti, C., Li, L. and Vogelstein, B., 2017. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science, 355(6331), pp.1330-1334.
70. Vanitha, C.D.A., Devaraj, D. and Venkatesulu, M., 2015. Gene expression data classification using support vector machine and mutual information-based gene selection. procedia computer science, 47, pp.13-21.
71. Van't Veer, L.J. and Bernards, R., 2008. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, *452*(7187), p.564.
72. Vapnik, V., 1998. Statistical learning theory. 1998. Wiley, New York

73. Vergara, J.R. and Estévez, P.A., 2014. A review of feature selection methods based on mutual information. Neural computing and applications, 24(1), pp.175-186.
74. Ware, B.R., McVay, M., Sunada, W.Y. and Khetani, S.R., 2017. Exploring chronic drug effects on microengineered human liver cultures using global gene expression profiling. *Toxicological Sciences*, *157*(2), pp.387-398.
75. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M. and Cancer Genome Atlas Research Network, 2013. The cancer genome atlas pan-cancer analysis project. Nature genetics, 45(10), p.1113.
76. Yilancioglu, K. and Konuk, M., 2015. Classification of Schizophrenia Patients by Using Genomic Data: A Data Mining Approach. The Journal of Neurobehavioral Sciences, 2(3), pp.102-104.
77. Zhang, H., Xie, Z., Yang, Y., Zhao, Y., Zhang, B. and Fang, J., 2017. The correlation-base-selection algorithm for diagnostic schizophrenia based on blood-based gene expression signatures. BioMed research international, 2017.
78. Zhang, L., Yang, T., Yi, J., Jin, R. and Zhou, Z.H., 2016, February. Stochastic Optimization for Kernel PCA. In AAAI(pp. 2315-2322).
79. Zhao, G. and Wu, Y., 2016. Feature subset selection for cancer classification using weight local modularity. Scientific reports, 6, p.34759.
80. Zhong, W., 2014. Feature selection for cancer classification using microarray gene expression data (Doctoral dissertation, University of Calgary).
81. Zhou, X. and Belkin, M., 2014. Semi-supervised learning. In Academic Press Library in Signal Processing (Vol. 1, pp. 1239-1269). Elsevier

Mr. S. Karthik received M. Tech in Software Engineering from Vellore Institute of Technology, Vellore, India in 2017. He is currently full time Research Scholar pursuing PhD in Information Technology & Engineering from VIT.His current research interests are in Bioinformatics, Machine Learning and Computational Intelligence.

Dr. M. Sudha, is currently as Associate Professor in the department of Information Technology of School of Information Technology and Engineering at VIT. She has 17 years of teaching cum research experience. Her research expertize are in the field of Medical Informatics, Machine Learning and Ambient Intelligent Computing.