Deep Learning Classifier for Gene Expression Datasets using a Hybrid LSTM Network

Immaculate Mercy A, Chidambaram M

Abstract*: A deep learning system Long Short-term memory (LSTM) is incorporated for the classification of differentially expressed genes which causes certain abnormalities in the human body. The LSTM is employed along with the K-Nearest Neighbour (KNN) algorithm so as to achieve the classification to its precision. The feature selection process plays a vital as some of the existing algorithms tend to neglect the features of concern. The classification further leads to enhanced prediction method. The K-Nearest Neighbour method is used to filter the correlation degree between each value with target value. This hybrid algorithm has a clear leverage over the existing methods. This work is well supported by the Feature Selection which includes a hybrid of Principal Component Analysis and the CHI square test. This hybrid approach provides with a good feature selection which aides in the seamless flow of the process towards classification and prediction. The Eigen values and the Eigen vectors are computed which effectively leads to the identification of Principal components. The Chi Square test is implemented for calculating the scores. The features that are obtained are ranked by these scores and the datasets which has the highest scores are further taken for training. The algorithms employed in this work has a clear advantage over the Bayesian networks as the Bayesian networks are prone to errors within the layers which may cause the values to explode or vanish. The accuracy of the classification and the prediction process achieved is unsurpassed when compared to the existing methods.

***Keywords* : Deep learning, LSTM, KNN, PCA, CHI square**

I. INTRODUCTION

Machine learning has become one of the most needed trends in every industry in the world. With the inception of industrial revolution 4.0 around the globe the industries are craving for better utilization and profits by the application of these technologies preferably Machine learning algorithms. The impact of these technologies has gone beyond the internet industry. It is of profound use in automobile, legal, agriculture, health industries and service sectors. Novel machine learning models are powered by neural networks. It is be neural networks are deployed at the edge layer. This leads to powerful predictive analysis. The amount of data that is collected everyday in any industry is massive. These data needs appropriate analysis and interpretation methods for use in decision making.

Emerging machine learning trends provide the means to collect, store and analyze these data. As the biological data is on the rise a lot of complex biological problem needs to be solved by the use of suitable algorithms. Predominantly the gene expression datasets are in need of better predictive methods which would be of great help to the health sectors.

In deep learning each level learns to transform its input data into slightly more abstract and composite representation. A deep learning process can learn which features to place at which level on its own. They have a substantial Credit Assignment Path (CAP) depth. The CAP is the chain of transformations from input to output. CAPs describe potentially causal connectors between input and output .Deep models are able to extract better features than shallow models and hence aids in learning he features effectively.

II. RELATED WORKS

The signature extraction [1] and Meta-Analysis of gene expression data compared the performances of various classifiers which employed a 10-fold cross validation technique. The feature selection that has been employed was time consuming. The sample clustering and group pairing methods separate the samples into small groups and are merged in a hierarchical pattern. Since only some series of the clusters are considered for enumerating does not lead to useless pairs. These pairs that have been obtained were not further considered for prediction method. The Gene Co-Expression Networks [2] used the Pearson correlation function to compare and measure the distance between a gene pair by using function adjacency. An undirected DCN was constructed and the analysis were performed on various parameters like degree, topological coefficients, clustering ,average no. of neighbors etc., The data that has been handled was microarray samples of gene data. As the samples were only few the precision of the method employed suffers from scalability and probabilistic measures. The Classification of Genes using Recurrent Neural Networks [3] used the tissue data set expression data. The functional differences between various rarely expressed genes were taken as samples. It employed supervised classifier recurrent neural network to discriminate features for classifying the genes. This method provides way for facilitating the identification and expression of Genes. This work is limited to the way of working with gene expression datasets which forms the basis for all analysis and prediction on the genes. The Sparse PCA for Tumor classification [4] used a supervised discriminative sparse PCA. This method incorporates the discriminative information and sparsity into the model. The classifications are empirically verified through abundant, reasonable results.

Revised Manuscript Received on February 28, 2020.

* Correspondence Author

Immaculate Mercy A.*, Pursuing, Ph.D, Computer Science, Bharathidasan University. Tamil Nadu, India.

Chidamabaram M., Assistant Professor, Computer Science, Rajah Serfoji Government College (RSGC), Thanjavur. India.

The methods that were utilized were not able to capture all the structure of the data.

The robustness of this method needs to improvised for further practical applications. Knowledge elicitation using PCA, Pearsons Chi Square [5] was able to learn the data using the existing method such as Naïve Bayes, Bayesian networks and Multinomial Logistic regression.

The emphasis has not been on feature selection as it is the feature selection which paves path for accuracy in learning and prediction.

III. METHODOLOGY

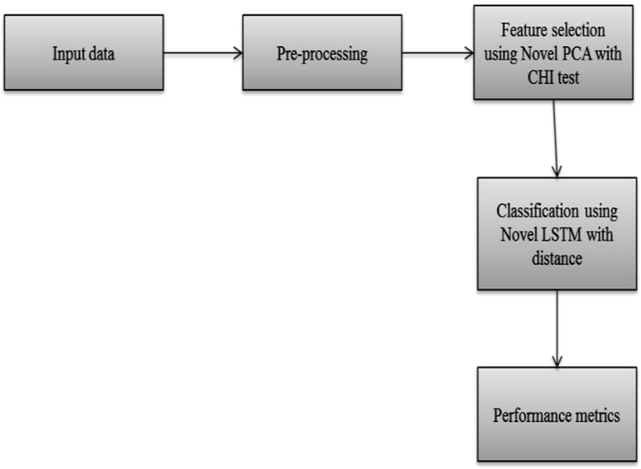
Input data
Pre-processing
Feature selection using Novel PCA with CHI test
Classification using Novel LSTM with distance
Performance metrics

Figure 1 Proposed Work Flow

The Proposed work depicts the learning process by the use of LSTM and KNN which leads to classification and the prediction process. The input data goes into the preprocessing stage where the irrelevant and noisy data are removed. After the preprocessing the data is fed for feature selection method which uses the Novel PCA with CHI test. Once the features are extracted the classification process proceeds by the way of using the Novel LSTM with the distance which is achieved by using the KNN method

A. Feature selection using Novel PCA with CHI test

The data after the preprocessing goes for the feature selection. The respective structure of the data is found and its is converted into a Matrix designated as M_{ij} . The matrix is then normalized for the appropriate values which in turn yields a normalized matrix N_{ij} . The covariance between the elements is found. The Eigen values and Eigen vectors are calculated for each of the arrived covariance values.

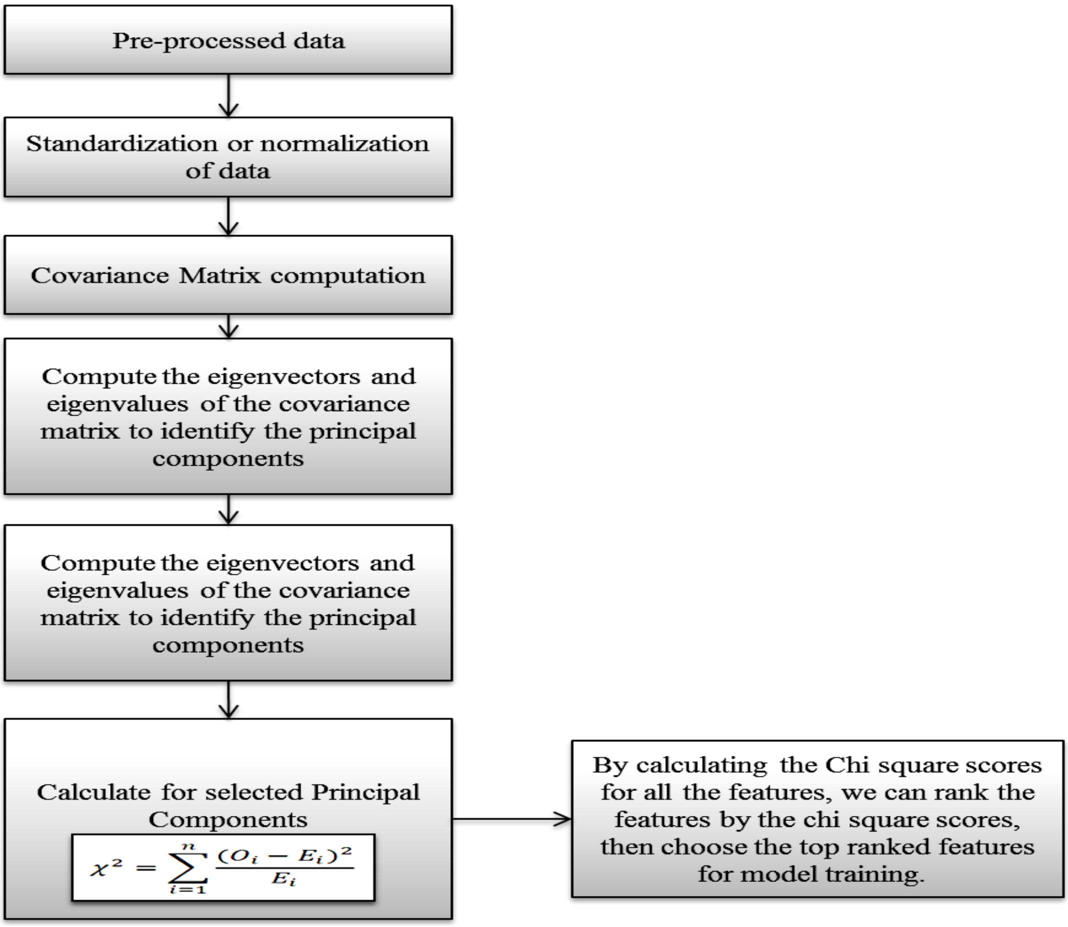
Pre-processed data
Standardization or normalization of data
Covariance Matrix computation
Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
Calculate for selected Principal Components
$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$
By calculating the Chi square scores for all the features, we can rank the features by the chi square scores, then choose the top ranked features for model training.

Figure 2 Feature Selection Process

The computed Eigen vectors and Eigen Values are found to identify the Principal components. CHI Square scores are calculated for all the features. The new features are calculated which is given by N_{ij}*. Analyzing the newly arrived features we then try to identify the features with highest scores and the ranking is done. Only these features are considered for the study and the rest are dropped. These arrived features are used for the training process. This step has been of concern as sometimes we may arrive at a classification which may not be of chosen interest or the domain of study.

Algorithm:PCA with CHI

Step 1: Input dataset
Step 2: find the structure covert into matrix M_{ij}

Step 3: Standardize or normalize the matrix $N_{ij} \leftarrow M_{ij}$
Step 4: Find Covariance of N_{ij} $by N_{ij} \leftarrow N_{ij}$ T N_{ij}
Step 5: Calculate Eigen Vectors and Eigen Values for N_{ij} with λ
Step 6: Sort the Eigen Vectors$Eigen\ values = \lambda_1, \lambda_2, \lambda_3$ ….. λ_n and $Eigen\ vectors = P$
Step 7: Calculate the new features which is$N_{ij}{}^*$
Step 8: Drop unimportant features from the $N_{ij}{}^*$
Step 9: Find chi square for$N_{ij}{}^*$ by the formula

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

x^2 = Pearson's cumulative test statistic, which asymptotically approaches a x^2 distribution. O_i= an observed frequency; E_i = an expected frequency, affirmed by the null hypothesis; n = the number of cells in the table.
Step 10: Take selected feature from the chi square step

B. Classification and Prediction using LSTM and KNN

The features that have been extracted from the previous step are sent for the learning process. As the LSTM comprises of cell, input gate, output gate and a forget gate it has the capability of remembering the values achieved in the previous stages. The weights are initialized for each of the neurons. The training samples are read and the distance between them is computed. Now the samples are trained to arrive at the required output. The KNN model is used to filter the correlation degree between each value with a specified target value. These values that are arrived are then used for the training and prediction.

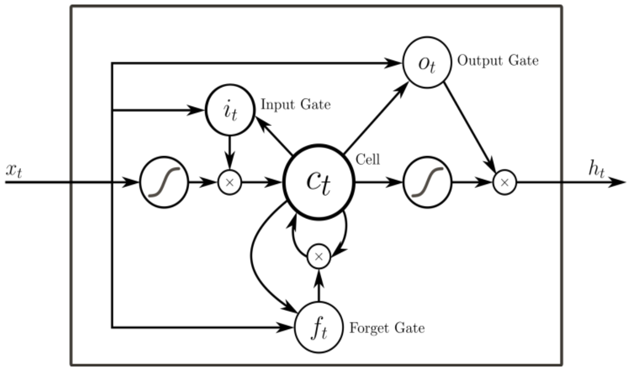

Figure 3 A Normal LSTM Network

In this process, the weights are initialized for each neurons , that is the features of interest for the study. The training samples are then fed into the system. After they are read, the distances between each of the samples are computed. This computation result if further considered for a comparison to see if it lies within the scope of the expected or the targeted output. Various parameters are used for the study.

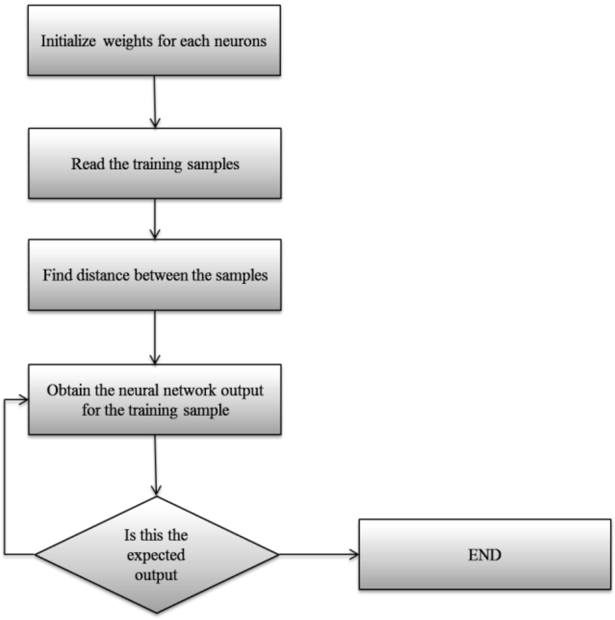
Initialize weights for each neurons
Read the training samples
Find distance between the samples
Obtain the neural network output for the training sample
Is this the expected output
END

Figure 4 NOVEL LSTM with KNN

The algorithm used for this purpose is the Novel LSTM with KNN.

Algorithm: Novel LSTM with KNN

Step 1. Calculate the Euclidean distance between adjacent gene values with the test values according to the below formula

d=$\sqrt{\sum_{i=0}^{q}(k_{1i} - k_{2i})^2}$

Step 2. K is the value selected by KNN
Step 3. Select mostly related value with the test value.
Step 4. Predict Gene behaviour with LSTM network, respectively, in selected values according to the below formula.
$\hat{X} = \Phi([x_1\ \ x_2\ \ldots x_{E-1}\ \ x_E])$
Step 5. Calculate the RMSE for the predicted value.
Step 6. Update the weights.
Step 7. Repeat Steps 2-6 with the different K.
Step 8. Find the smallest RMSE in all the different K.
Step 9. Obtain the predicted value in the test data when RMSE is the smallest.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

The Euclidean distance between each of the adjacent gene is calculated with the test values using the formula

$$d=\sqrt{\sum_{i=0}^{q}(k_{1i} - k_{2i})^2}$$

The K value needed for the distance calculation is achieved by KNN. The values are selected in such a way that the most related value is closest to the test value. The behavior of the gene in the LSTM network is selected using the formula$\hat{X} = \Phi([x_1 \;\; x_2 \; \dots x_{E-1} \;\; x_E])$. 'Φ' is the LSTM model. The RMSE (Root Mean Square Error) is calculated for each of the predicted values. The Weights of each neuron are updated based on the predicted values. Likewise the predictions are found for different K values. The smallest or the negligible RMSE values for different K values are found. The K values which has the smallest RMSE value is hence taken for the prediction and the genes are predicted.

IV. RESULTS AND DISCUSSIONS

The Classification and prediction accuracy of the proposed methods are achieved by the ways of experimental results. The genes which have the abnormalities are clearly predicted with the aid of the employed modified and converged algorithms. The proposed method has a clear advantage over all the existing methods especially with the nature of data that has been considered for the study.

Data set :

The dataset taken for the study is the GSM data which is gene expression dataset taken from the GEO database from the NCBI repository. These GEO datasets are curated datasets. This database stores curated gene expression Datasets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Data set records contain additional resources including cluster tools and differential expression queries. The GSM data simulates the human metabolism that this feature has enabled in systematic evaluation of metabolic features of human disease. The analysis is highly reliable for further treatment and predictions. The datasets that has been taken for the study as the test datasets are the lung disease datasets and the training datasets are the unknown diseased or normal datasets on which the learning process is applied to go for classification and further predictions. The analysis is highly reliable for further treatment and predictions. 13 features are selected and the data are filtered into respective features and the labels are assigned for the values and the class data are arrived for all the featured datasets.

Table 1: Dataset for the Analysis

DATASET	GDS4930
Dataset_platform_organism	Homo sapiens
Dataset_title	Fetal lung development
Dataset_sample_type	RNA
Dataset_sample_count	38
Dataset_reference_series	GSE14334
IDENTIFIER	GSM358668
	GSM358657
	GSM358633
	GSM358634
	GSM358638
	GSM358656
	GSM358631
	GSM358637
	GSM358650
	GSM358667
	GSM358654
	GSM358660
	GSM358652
	GSM358651
	GSM358665
	GSM358666
	GSM358658
	GSM358655
	GSM358662
	GSM358636
	GSM358639
	GSM358635
	GSM358640
	GSM358663
	GSM358632
	GSM358661
	GSM358653
	GSM358664
	GSM358659
	GSM358645
	GSM358644
	GSM358646
	GSM358648
	GSM358649

The Process proceeds as follows. We decompose the genome-wide expression patterns in 38 embryonic human lungs (53-154 days post conception/dpc) into their independent, dominant directions of transcriptomic sample variation in order to gain global insight of the developing human lung transcriptome. The characteristic genes and their corresponding bio–ontologic attribute profile for the latter were identified. We noted the over–representation of lung specific attributes (e.g., surfactant proteins) traditionally associated with later developmental stages, and highly ranked attributes (e.g., chemokine–immunologic processes) not previously reported nor immediately apparent in an early lung development context.

We defined the 3,223–gene union of the characteristic genes of the 3 most dominant sources of variation as the developing lung characteristic sub–transcriptome (DLCS). It may be regarded as the minimal gene set describing the essential biology of this process. The developing lung series in this transcriptomic variation perspective form a contiguous trajectory with critical time points that both correlate with the 2 traditional morphologic stages overlapping -154 dpc and suggest the existence of 2 novel phases within the pseudoglandular stage.

To demonstrate that this characterization is robust, we showed that the model could be used to estimate the gestational age of independent human lung tissue samples with a median absolute error of 5 days, based on the DLCS of their lung profile alone. Repeating this procedure on the homologous transcriptome profiles of developing mouse lung 14–19 dpc, we were able to recover their correct developmental chronology. Whole human fetal lung gene expression profiling from estimated gestational ages 53 to 154 days post conception.

The graph for the above set of series of data with their labels and their counts in thousands are given as below:

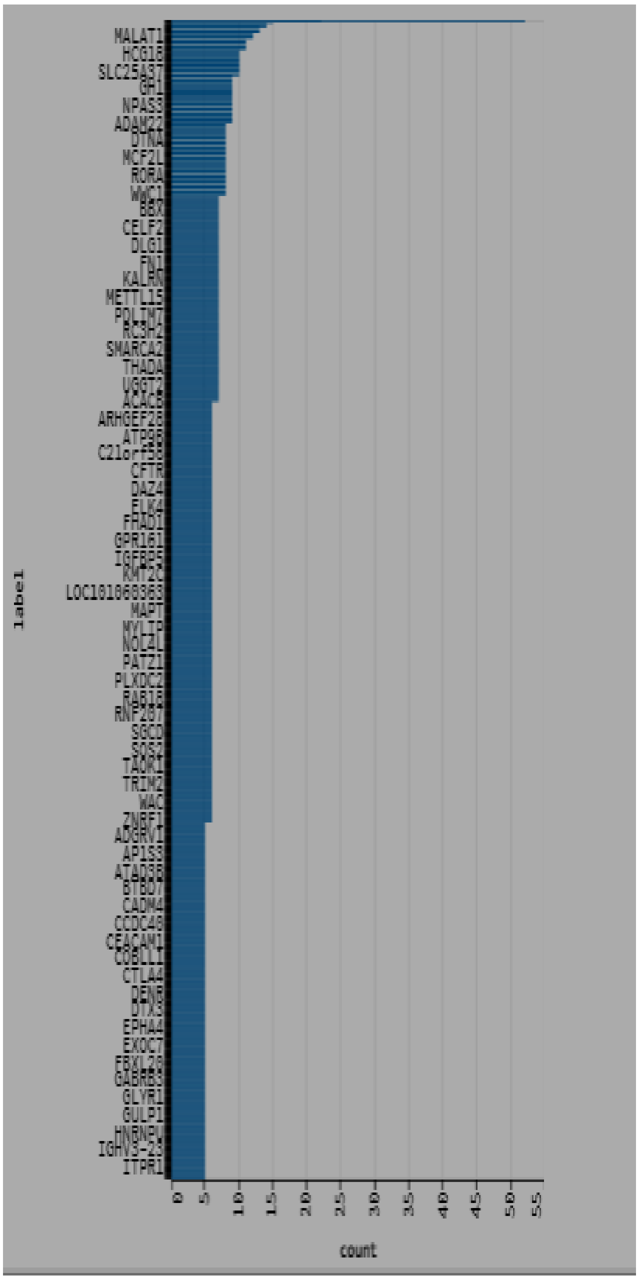

Figure 5 Labels Taken for the study

The other set of data included in this study is the adult and fetal datasets which has been curated and considered for further classification process.

RNA from circulating blood reticulocytes was utilized to provide a robust description of genes transcribed at the final stages of erythroblast maturation. After depletion of leukocytes and platelets, Affymetrix HG-U133 arrays were hybridized with probe from total RNA isolated from blood sampled from 14 umbilical cords and 14 healthy adult humans. Keywords: Adult vs. fetal reticulocyte transcriptome comparison Peripheral blood from 14 normal adult and 14 human umbilical cord were purified to obtain packed red blood cells and their purity was assessed by CELL-DYN4000. The contaminated WBC count was less than 1 out of 1 million cells. Total RNA from purified reticulocytes included in the purified red blood cells was extracted, labeled, and hybridized onto Affymetrix HG-U133 A and B arrays. These datasets have been included since they are in the domain of the study and awaiting the predictions.

The dataset description is as follows:

Table 2: Dataset for the study

DATASET	GDS2655
Dataset_platform_organism	Homo sapiens
Dataset_title	Fetal and adult reticulocytes (HG-U133A)
Dataset_sample_type	RNA
Dataset_sample_count	28
Dataset_reference_series	GSE6236
IDENTIFIER	GSM143586
	GSM143587
	GSM143588
	GSM143589
	GSM143590
	GSM143591
	GSM143592
	GSM143593
	GSM143594
	GSM143595
	GSM143596
	GSM143597
	GSM143598
	GSM143599
	GSM143572
	GSM143573
	GSM143574
	GSM143575
	GSM143576
	GSM143577
	GSM143578
	GSM143579
	GSM143580
	GSM143581

The graph for the above set of values and their respective labels are portrayed as follows:

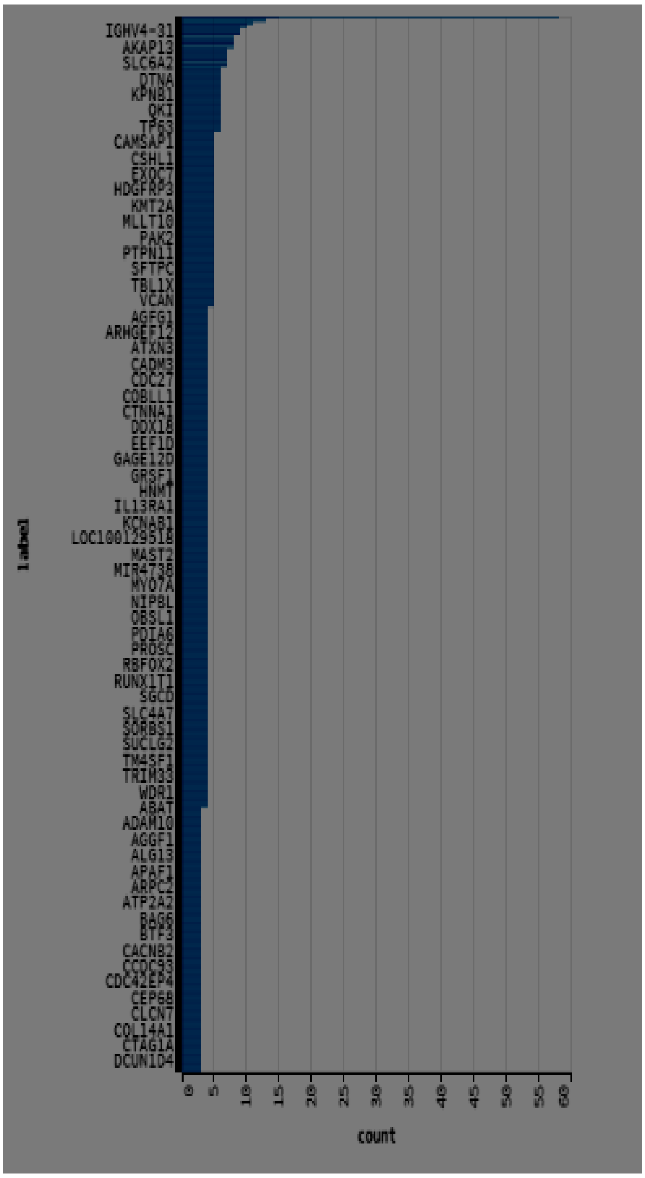

Figure 6 Labels with their counts in thousands

The confusion matrix has been arrived. The True positive and false positive values that have been arrived clearly shows that the proposed algorithm works better in comparison to the existing methods. These true positive and false positives have been computed for both the input datasets considered for the study. They are as follows:

Table 3
Fetal and adult reticulocytes (HG-U133A)

RMSE	0.024348
Accuracy	0.945652
Specificity	0.879218
Precision	0.944535
Recall	0.977215
F1-Score	0.960597
False Positive Rate	0.120782
False Negative Rate	0.022785
Negative Predicted Value	0.948276
Matthew's Correlation Coefficient	0.874433

The above Table 3 gives us a clear picture of the accuracy measures that has been carefully harnessed. The RMSE value when noted is very negligible which suggests that the proposed method gives us a clear cutting edge method over the other existing techniques.

A graphical view of the above table gives us clarity and it is quite obvious that this method could be suggested in the future.

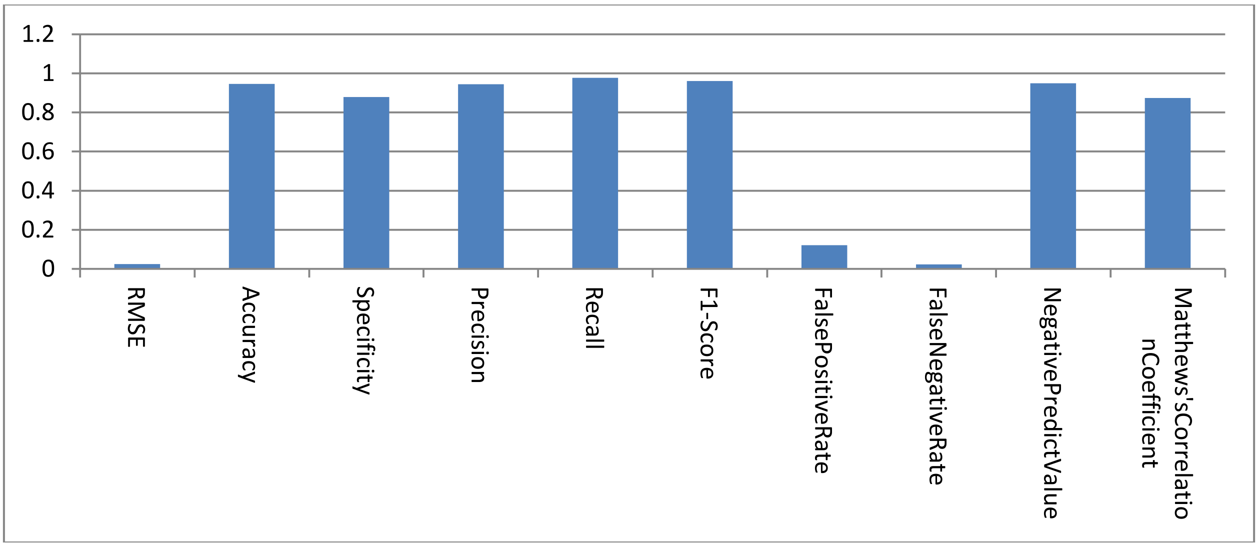

Figure 7 Accuracy Measures

The Confusion Matrix that has been obtained for the above datasets are as follows:

Table 4: Confusion matrix1

TP	1158
FP	68
FN	27
TN	495

It is observed from Table 4 that the True positives and the true negatives have been achieved to its maximum. The Euclidean distance between the adjacent genes is found. From these values the most closely related values are selected along with the test value. The gene behavior is predicted with the LSTM network for the selected values. The RMSE value that has been obtained is 2.43% which clearly depicts that this proposed algorithm has a clear cut over the existing algorithms. Now the test data are converted into a matrix which is further normalized. The covariance of the normalized matrix is obtained and the Eigen values ‘λ’ is obtained. These Eigen values are then sorted and stored in a vector. The new features are calculated for the obtained vector. The Pearson’s Cumulative Test Statistic ($\varkappa^2$) which asymptotically approaches a $\varkappa^2$ distribution where O_i is the observed frequency and E_i the expected frequency are calculated. The differences that have been obtained between the Oi and Ei is negligible. This additionally proves that the method proposed paves path for better accuracy and precision. The Accuracy that has been achieved is 94.56% . The precision over the classification is arrived at 94.45%.

The parameters have been tested using **H2O flow** which gives clear statistical evidence. H2O Flow, allows us to capture, rerun, annotate, present, and share the workflow. H2O Flow allows us to use H2O interactively to import files, build models, and iteratively improve them. Based on the models, we can make predictions and add rich text to create vignettes of the work. Flow’s hybrid user interface seamlessly blends command-line computing with a modern graphical user interface. However, rather than displaying output as plain text, Flow provides a point-and-click user interface for every H2O operation. It allows you to access any H2O object in the form of well-organized tabular data.

The training and the validation deviance graph obtained using H2O Flow

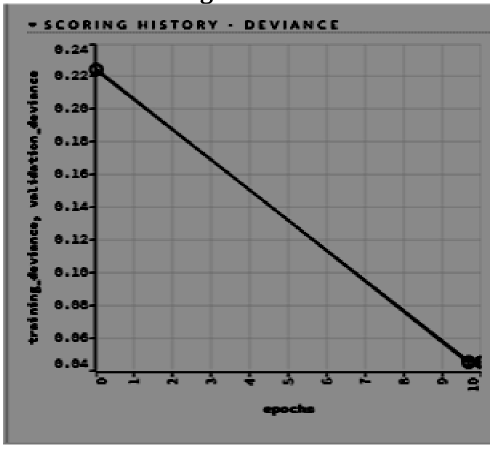

Figure 8 The Training and the Validation deviance graph

From the above graph it can be observed that the deviation decreases along the path. Likewise the accuracy measures over the fetal lung development datasets have been obtained as follows:

Table 5: Fetal Lung development accuracy measures

RMSE	0.044297189
Accuracy	0.925702811
Specificity	0.850574713
Precision	0.9675
Recall	0.941605839
F1-Score	0.954377312
False Positive Rate	0.149425287
False Negative Rate	0.058394161
Negative Predict Value	0.755102041
Matthews's Correlation Coefficient	0.756591887

The above Table 5 also harnesses the fact that the proposed algorithm has good RMSE value which is once again negligible. Similarly the accuracy, specificity, Precision, Recall, F1 score , Matthew’s correlation coefficient has worked to an acceptable set of values. The accuracy measures over various algorithms and the proposed method are tabulated as follows:

Table 6: Accuracy Measures Compared across various algorithms

Algorithm	Accuracy
RNN	0.899
RF	0.905
LDA	0.754
SDA	0.79
PCA	0.88
SPCArt	0.9
SDSPCA	0.91
Proposed	0.945

Analyzing the above table of facts various algorithms has given its performances based on the methods employed. The **RNN** algorithm is a very popular algorithm in deep Learning but still it lacks behind in the accuracy measures the is was able to give only 89.9%. Likewise The **RF** algorithm commonly known as the Random Forest method or Random Decision forests are ensemble learning methods. The selection features employed here fails to take the features of interest considered for the study. The **LDA** (Latent Dirichlet Allocation) is a generative statistical method that allows set of observations to be explained by unobserved groups which explains why some data are similar. Since this algorithm makes use of Bayesian inference model the accuracy and the dimensionality curses are not taken care of well. The **PCA** (Prinicpal Component Analysis) when employed worked out to only 88%. Similarly the **SPCArt** and the **SDSPCA** doesn’t have a clear edge over the set of data considered for the study.

These are clearly depicted in the following chart:

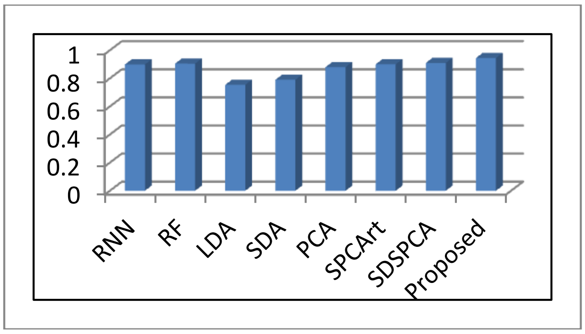

Figure 9 Comapartive Accuracy measures over Various Algorithms

Now considering the RMSE values a comparative measure has been worked taking the LSTM network and the proposed Algorithm.

Algorithm	RMSE
LSTM	0.2
Proposed	0.04

Figure 10 RMSE values Comparison

The LSTM when used separately the RSME works out to 0.2, nut in the proposed methodology we used a hybrid of LSTM with KNN which gives an RSME value of 0.04 which is a negligible error rate.

Considering the running time of the algorithm it has been evaluated for various algorithms which are as follows:

Algorithm	Time
LDA	0.23
SDSPCA	0.52
PCA	0.18
PathSPCA	0.15
Proposed	0.14

Figure 11 Running time comparison

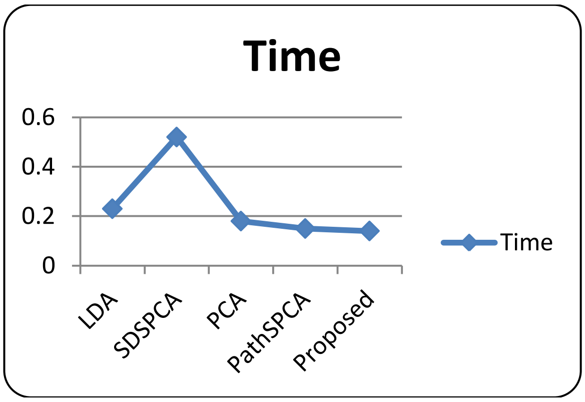

Figure 12 Running Time Comparison

From all the above statistical measures we could see that all the values that have been obtained for testing the classification accuracy, the deviation values, the recall, the specificity are all satisfactory. It could also be suggested that this method has a clear advantage over the existing methods.

A total of 54,395 datasets have been considered for the study. Out of which more than 22,500 datasets are seen to have abnormal features. Inferentially these 22,500 datasets or genes or persons are likely to have a lung disease which varies in degrees of occurrence.

The following graph gives the Normal and the abnormality measure of the datasets.

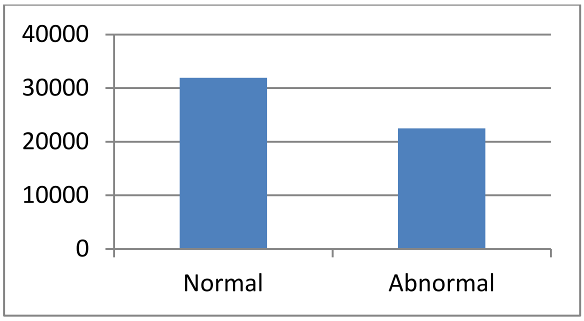

Figure 13 Prediction

V. CONCLUSION

The gene classification using the LSTM network along with the Hybrid PCA and KNN algorithms have been effectively used for the classification and the prediction process. Considering the unstructured format of the GSM GEO data that has been taken a careful data preprocessing has been taken care of by the LSTM network. The required features are selected from the datasets and the covariance and the Eigen values are obtained. The values that have been achieved clearly prove the effective use of the algorithm. The accuracy value that has been obtained clearly separates the normal and the abnormal datasets. As this works for the lung disease the same methods can also be recommended for all kinds of severities that occur in the human body. When considering for further works the explosive amount of data will not have a deteriorating effect on the process which has been guaranteed.

REFERENCES

1. Shankai Yan and Ka-Chun Wong, "GESgnExt: Gene Expression Signature Extraction and Meta-analysis on Gene Expression Omnibus" in Journal of Biomedical and Health Informatics vol. 24, issue 1, January 2020
2. Su-Ping Deng,Liu Zhu and De-Shuang Huang ,"Predicting Hub Genes Associated with Cervical Cancer through Gene Co-Expression Networks " *IEEE Transactions on Computational Biology and Bioinformatics*, Vol 13, No.1, January 2016.
3. Lei Chen, Xiao Yong Pan, Yu-Hang Zhang, Min Liu, Tao Huang, Yu-Dong Cai , "Classification of widely and rarely expressed genes with recurrent neural network" *Computational and Structural Biotechnology Journal* December 2018.
4. Xing-Lin Hsu, Po-Yu Huang and Dung-Tsa Chen , "Sparse PCA for Cancer Calssification " in *PMC – US National Library of Medicine, National Institute of Health,* Vol 3, June 2014.
5. P.Paokanta, "β- Thalassemia Knowledge Elicitation Using Data Engineering: PCA , Pearson's Chi square and Machine Learning " *International Journal of Computer Theory and Engineering* Vol 4, No.5, October 2012.
6. D. S. Huang and C. H. Zheng, "Independent component analysis based penalized discriminate method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, no. 15,
7. World Health Organization, *World Cancer Report 2014*, pp, Chapter 5.12, 2014.

8. Mahmoud Ahmed Ismail Shoman, Reda El-Khoribi, “Gene Expression based Cancer classification “ *Egyptian Informatics Journal* Vol 18(3), December 2016.
9. Ayyad SM, Saleh Al, Labib LM, “Gene Expression Cancer Classification using modified K-Nearest Neighbors Technique”, *Biosystems, Science Direct , Elsevier* Vol 176, February 2019.
10. Manish Babu, Kamal Sarkar, “A Comparative study on Gene *selection Methods for Cancer classification using microarray data” 2016 International Conference on Research in Computational intelligence and Communication Networks , IEEE,* Jan 2017
11. Zhenfang Hu, Yueming Wang, Gang Pan, Zhaohu Wu*, “Sparse Principal Component Analysis via Rotation and Truncation”*, Cornell University March 2014.
12. Chun-Mei Feng, Yang Xu, Jin-Xing Liu, Xing-Lian Gao, Chun-Hou-Zheng, *“Supervised Discriminative Sparse PCA for Characteristic Gene Selection and Tumor Classification on Multiview Biological Data”,* IEEE Volume 30 Issue 10 , October 2019.

AUTHORS PROFILE

Immaculate Mercy .A, is pursuing her Ph.D in Computer Science from Bharathidasan University. She holds a Masters Degree in Computer Applications, Masters in Computer Science Engineering and M.Phil in Computer Science. She has over 16 years of experience in teaching at the Under Graduate and the Post Graduate level and 8 years of industry experience, which involved in software development and e-content writing. She also has a proven experience in Corporate Training. Her research work focuses on Data Science, Prediction Analysis, Data Mining, Big Data Analytics and Cloud Computing.

Chidamabaram .M, pursued his Masters in Computer Science from Bharathidasan University, M.Phil from Bharathidasan University, M.B.A from Periyar University and Ph.D from Vinayaka Mission’s. in Computer Science. He has over 21 years of experience in teaching at the undergraduate and post Graduate level. He has guided over 25 scholars towards M.Phil degree and 8 scholars towards Ph.D degree. He has published more than 40 research papers in various National and International journals, 8 conference papers .He is currently working as an Assistant professor in Computer science in RSGC, Thanjavur. His areas of Research are Cloud Computing, Grid Computing, Data Mining and Big Data Analytics.