Molecular Classification of Cancer by Gene Expression Monitoring

Shivangi Goswami, Sanjay Katta
Department of Computer Science
Michigan Technological University
Houghton, MI, USA

I. BACKGROUND

At the Chromosomal level, the human body carries genetic information through deoxyribonucleic acid (DNA). The bases that code the DNA molecule are namely Adenine, Guanine, Cytosine ,and Thymine. These DNA molecules form the basis of life through which the genetic information is carried down to the offspring.

The DNA makes its copy of itself through a biological process of transcription. These DNA copies undergo another process known as reverse transcription, through which transcripts of ribonucleic acid (RNA) are formed. RNA molecules play a critical role in gene expression[1] by combining through different regulatory sequences or motifs and synthesizing into protein by the process of transcription. Therefore these genes are responsible for encoding proteins which further are employed to direct the cell functionality. Mutations in these genes can lead to diseases like cancer. This paper focuses on the classification problem of gene alleles and has identified cancer into two main classes, namely acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).

II. MOTIVATION

The interactions between RNA and regulatory factors that bind them for mediating the protein synthesis are very crucial. Mutations in these genes can lead to life-threatening diseases such as cancer. There are many state-of-the-art techniques used to date for the identification of these genes, but these techniques are fairly time-consuming and expensive. In this paper we have used machine learning approaches to address the classification problems of genes on the basis of identified cancer classes.

The machine learning algorithms used in this project are logistic regression classification algorithm, KNN algorithm, Random Forest classification algorithm, Support vector machine and finally neural network. All these models are further analysed by calculating the accuracy.

III. ALGORITHM/METHODS

The algorithms and methods used in implementation of molecular classification of cancer by gene expression in this project are described below

K-Means Clustering - K-means clustering a s the name suggests is a clustering technique in which the entire data set is deduced to the small clusters to get the perception of what target class the group belongs to. The algorithm works on the following steps, first K centroids are randomly located in a plane or space, the next two steps are reiterative until the convergence is achieved, (i) for each point in the dataset, the algorithm finds the nearest centroid (the distance between centroid and point can be find by using different distance matrices like euclidean distance or the manhattan distance) and then the point is assigned to that centroid (ii) for each cluster i ranging from 1 to n, a new centroid which is essentially the mean of all point assigned to cluster i in preceding step.

Naive Bayes Algorithm - Naive Bayes theorem, as the name suggests forms its basis from the Bayes theorem for conditional probability with a naive assumption that the features or the attributes are not correlated or correspond to each other and tries to find the

conditional probability of the target variable given the probabilities of features.

Logistic Regression - Logistic regression which can be considered as somewhat similar to linear regression, where linear regression explores the continuous association between an independent variable ,let's say *Q* and dependent variable ,let's say *R*, logistic regression predicts whether something is correct or not correct (0 or 1), like, whether the cancer is of AML type of ALL type. Logistic regression, rather than considering to fit a straight line across the data, fits into an “S” shaped ” logistic or sigmoidal function”.
The curve goes from zero to one which means that the curve tells you the probability of whether the gene is AML or ALL.

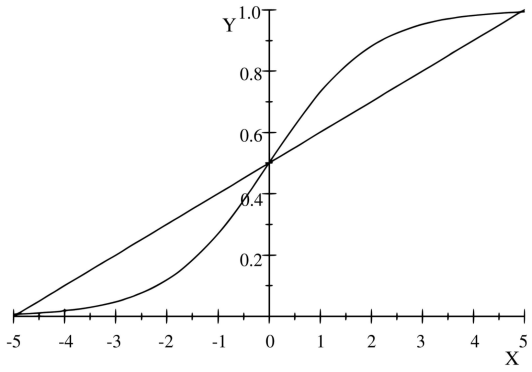

Figure 1 : Plot depicting linear and logistic regression. [4]

Support Vector Machine - Another algorithm used in this paper to classify the cancer type, is the support vector machine which is a supervised learning algorithm. In SVM to classify the data, a threshold value is chosen and when we have a new observation, it is classified based on the threshold value. When the data is three-dimensional, the Support Vector Classifier forms a plane instead of a line[3]. and we classify new observations by determining which side of the plan they are on.

Random forest algorithm - Random forest algorithm is one of the supervised machine learning algorithms for a classification as well as for regression problems. It creates the forest with several decision trees, in commonality, if there are a greater number of trees in the forest the accuracy and the predicted rate increases. In a random forest several decision trees are joined together contrary to a single decision tree to classify a new observation, each tree in the forest votes for a particular class whereas overall the forest selects the class for classification which has the highest number of votes.

Neural Network - There is huge growth of use of neural networks while addressing the biological problems like gene classification, binding sites predictions[5]. Neural networks work on the line of neurons adapted from human biology. As neurons in humans form the core of the central nervous system similarly neurons form the core processing unit in the neural network. The first layer which is the input layer obtains the input whose output is predicted by the final layer. The hidden layers are the main computational layers. The inputs fed to the input layer are multiplied or simply a dot product is taken with the assigned weights and the sum of all these are sent to the hidden layers (bias can also be added). These values go through an activation function also known as threshold function (example Relu and Sigmoid), the results decides the activation or deactivation of a particular neuron and this step continues over the next layers of network through which the data move forward also referred to as “forward propagation”, the neuron with higher values in the output layer determines the output. To reduce the error the model trains itself multiple times by updating the weights assigned to it and this is achieved by another process of back propagation.

IV. DATASET

The dataset that we have used in the project has been retrieved from the gene expression dataset evidently demonstrated in the study(Golub et al.) which was published in 1999[2]. The dataset

contains samples from bone marrow and peripheral blood and contains the quantification of the samples for acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The data has been retrieved from 78 patients with a total of 7200 attributes.

1) The Dataset has 3 files.
2) The file 'actual.csv' which tells about the patients who have ALL and AML.
3) The two other files have the information about gene expression and each file has about 7200 rows.
4) The dataset is open source.

V. DATA PREPROCESSING

The Gene expression dataset contains 7200 gene descriptions, which all constitute the different attributes of the dataset. To proceed with any sort analysis on this dataset, preprocessing it becomes essential. Firsty, we dropped the columns from the dataset which weren't required for the analysis and later to reduce the number of dimensions in the dataset principal component analysis (PCA) is performed to deduce the data to just features relevant for predictions.

VI. METRICS AND EVALUATION CRITERION

To evaluate the performance of our model, we will use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) where a comparison between the predicted rating and the true rating will be used to test the model, where MAE (Mean absolute error) and RMSE (Root Mean squared error) is given by,

$$\mathbf{MAE} = \frac{1}{|T|} \sum_{(u,i)\in T} |r_{ui} - \hat{r}_{ui}|$$

$$\mathbf{RMSE} = \sqrt{\frac{1}{|T|} \sum_{(u,i)\in T} (r_{ui} - \hat{r}_{ui})^2}$$

Where, r_{ui} is the predicted value for the movie i, $\hat{r}_{ui}$ is the observed true value for the movie i and T is the total number of rated movies.

VII. EXPERIMENTAL DESIGN

The dataset already contains different files for the train set and test set. After preprocessing of the data, the data was standardized. Figure 1 and figure 2 depicts the distribution of data before and after the features of the data were standardized. In figure 1 the data is more concentrated in one region and shows a peak close to zero, whereas after applying standardization the is more spread out and follows the gaussian distribution.

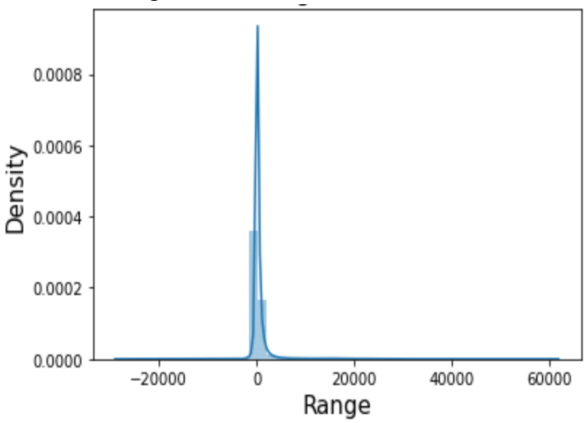

Figure 2 : Data distribution before standardization

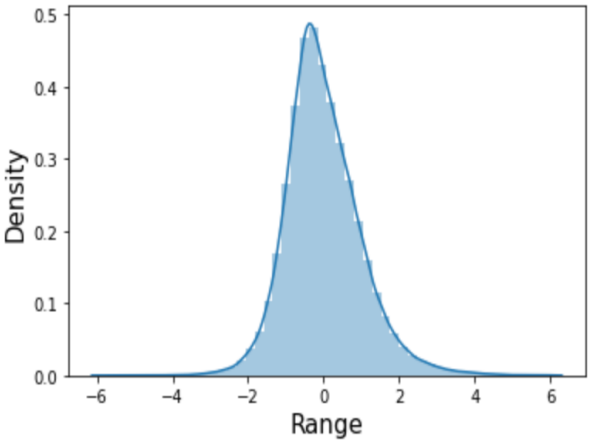

Figure 3 : Data distribution after standardization

Principal component analysis (PCA) was performed on the data for the dimensionality reduction, after which the train and the test data had 32 columns.

VIII. RESULTS

K-Means Clustering - The first algorithm that is considered in this paper is the K-means clustering algorithm. After training our model on 2 clusters we have evaluated the performance of our model using the MAE and RMSE metrics and finally also calculated the accuracy of the model. Table 1 shows the value of all three metrics for K-Means clustering.

RMSE	MAE	ACCURACY
0.7173	0.5145	0.7350

Table 1: Performance metrics for K Means clustering

The confusion matrix generated using K Means clustering shown in figure 4 depicts that out of 38 samples 17 ALL samples and 8 AML have been predicted correct.

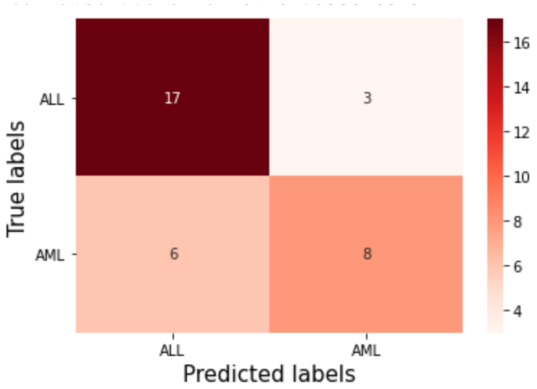

Figure 4: K-means Confusion matrix

Naive Bayes Algorithm - We used Gaussian Naive bayes to train and fit our model. It can be seen from table 2 that Naive bayes classification performed significantly better than the K - means clustering algorithm. This could also be because of the assumption that cancerous genes follow the gaussian distribution.

RMSE	MAE	ACCURACY
0.5450	0.0882	0.9120

Table 2: Performance metrics for Naive bayes

The confusion matrix generated using Naive bayes shown in figure 5 depicts that out of 38 samples 18 ALL samples and 13 AML have been predicted correct.

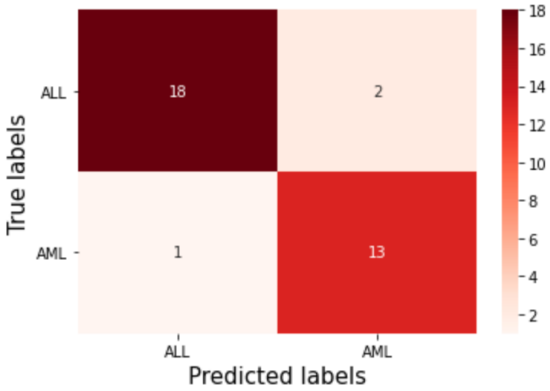

Figure 5: Naive Bayes Confusion Matrix

Logistic Regression - To train the model on model on logistic regression we have first used grid search cross validation to tune the hyperparameters and following the best parameters the predictions were made. Table 3 shows the value of all three metrics for logistic regression algorithm.

RMSE	MAE	ACCURACY
0.4141	0.0294	0.9710

Table 3: Performance metrics for logistic regression

The confusion matrix generated using logistic regression shown in figure 6 depicts that out of

38 samples 19 ALL samples and 14 AML have been predicted correct and only 1 AML sample have been predicted wrong. So far, logistic regression seems to outperform all the models.

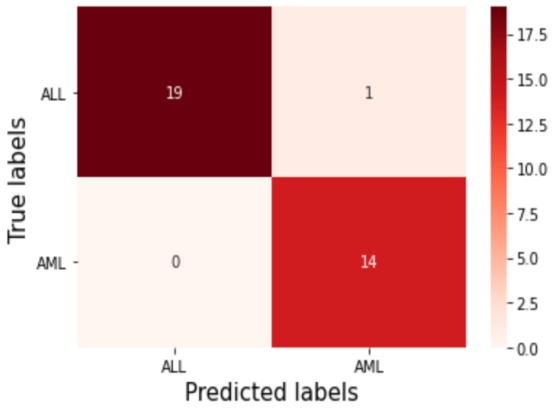

Figure 6: Logistic Regression Confusion Matrix

Support Vector Machine - To train the model on a support vector machine GridsearchCV using 3-fold cross validation was used to tune the hyperparameters and following the best parameters the predictions were made. Table 4 shows the value of all three metrics for the support vector machine.

RMSE	MAE	ACCURACY
0.5450	0.0882	0.9120

Table 4: Performance metrics for SVM

The confusion matrix generated using SVM shown in figure 7 depicts that out of 38 samples 20 ALL samples and 11 AML have been predicted correct.

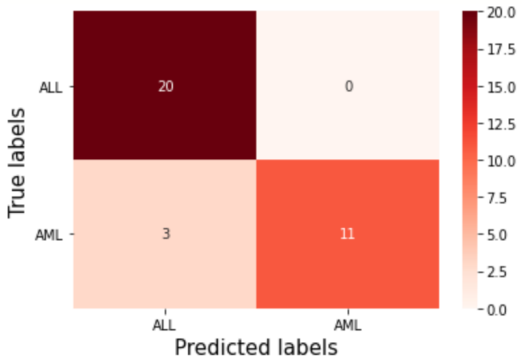

Figure 7: SVM Confusion Matrix

Random Forest Algorithm - After using the GridsearchCV using 3-fold cross validation to tune the hyperparameters, the predictions were made using the best parameters. Table 5 shows the value of all three metrics for the Random Forest.

RMSE	MAE	ACCURACY
0.6192	0.1471	0.8530

Table 5: Performance metrics for Random Forest

The confusion matrix generated using Random Forest shown in figure 8 depicts that out of 38 samples 20 ALL samples and 9 AML have been predicted correct.

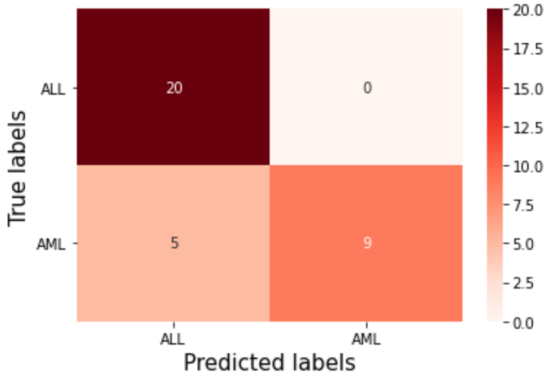

Figure 8: Random Forest Confusion Matrix

Neural Network - We have used keras to input the data into the convolutional layers. The first layer is the input layer which simply takes all the 32 samples as input. In the next convolutional layer the number 16 is because we didn't zero-pad the input samples before convolving with a rectified linear unit (ReLU) activation or threshold function. The last layer of the network uses the sigmoid activation or threshold function. The model was trained and validated using batch size of 16 and 50 epochs.

The accuracy achieved using the neural network is 0.7940 which is given in table 5.

RMSE	MAE	ACCURACY
0.6736	0.2059	0.7940

Table 6: Performance metrics for Neural Network

The confusion matrix generated using neural network shown in figure 9 depicts that out of 38 samples 19 ALL samples and 8 AML have been predicted correctly and a total of 7 samples have been misclassified.

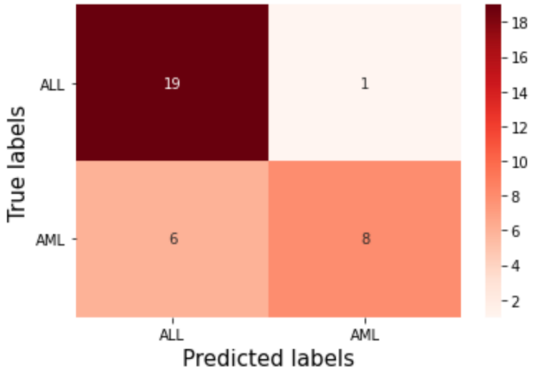

Figure 9: Neural network Confusion Matrix

IX. CONCLUSION

The predictive ability of the machine learning approach on the large volume of gene expression data has opened a myriad of opportunities to the diagnosis and treatments of various diseases based on gene expression profiling. To predict cancer morphologies, treatment success of patients and determine the relevant genes in the inference have gained favor in the recent studies. Therefore the classification of cancer becomes important as it also provides insights into possible treatment strategies. In this paper we have used six machine learning algorithms to classify the cancerous gene as ALL or AML namely, K-Means algorithm, Naive Bayes Algorithm, Logistic Regression, Support Vector Machine, Random Forest Algorithm and Neural Network.
From the performance metrics of all the above models logistic regressions seem to perform better than all the other models by correctly classifying 33 out 34 samples . The naive bayes and support vector machines also performed good with the accuracy of approximately 91%.

X. CONTRIBUTIONS

Our team consists of two members, to decide and work in this particular topic was a decision taken together, bearing in mind the growth of machine learning in the medicine field. To search for the appropriate dataset for the classification problem was well handled by Sanjay Katta, whereas the literature review and analysis regarding the cell biology and classification of disease was done by Shivangi Goswami. Both the members were equally involved in the development of the code as well as writing the paper.

XI. REFERENCES

[1] Denys V. Volgin, Chapter 17 - Gene Expression: Analysis and Quantitation,Editor(s): Ashish S. Verma, Anchal Singh, Animal Biotechnology, Academic Press,2014,Pages 307-325,ISBN 9780124160026.

[2]https://www.kaggle.com/crawford/gene-expression

[3] Brown, M., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M., & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America, 97 1*, 262-7 .

[4] A.A Fernandes, D.B Filho, E.C. Rocha, Revista de Sociologia e Política, Federal University of Pernambuco, Jan(2021)

[5] Y. J. Tan, K. S. Sim and F. F. Ting, "Breast cancer detection using convolutional neural networks for mammogram imaging system," 2017 International Conference on Robotics, Automation and Science (ICORAS), Melaka, Malaysia, 2017, pp.1-5.