Acute Leukemia Classification using machine learning algorithms and PCA

Mahalakshmi C.V., Aditya Gautam, Avik Bhattacharya, Ayush Vats, Nikhil Singh

[1]*Assistant Prof., C.S.E, Bangalore Institute of Technology*
[2]*student, B.E.(C.S.E.), Bangalore Institute of Technology*
[3] *student, B.E.(C.S.E.), Bangalore Institute of Technology*
[4]*student, B.E.(C.S.E.), Bangalore Institute of Technology*
[5]*student, B.E.(C.S.E.), Bangalore Institute of Technology*

Abstract – Acute leukemia is a proliferation of immature bone marrow-derived cells (blasts) that may also involve peripheral blood or solid organs. With the help of machine learning, patterns in genetic data can be found that were unknown to us earlier and these patterns can be very useful in making conclusions about diseases and disorders that are inherently genetic in nature.

This project focuses on gene expression analysis using Principal Component Analysis (PCA), a powerful technique for dimensionality reduction and pattern extraction in high-dimensional datasets. The objective is to uncover hidden relationships and identify key features that contribute to the variations in gene expression data. By applying PCA to a dataset consisting of gene expression measurements across multiple samples, we aim to gain insights into the underlying structure and patterns of gene expression.

Key Words: PCA, Acute Leukemia, machine learning, gene expression, classification, journals

1.INTRODUCTION *(Size 11, Times New roman)*

It highlights the need for effective techniques to analyze gene expression data and explores the limitations of traditional methods. PCA is then introduced as a valuable tool for dimensionality reduction, feature extraction, and data visualization. The underlying mathematical principles of PCA are explained, including the calculation of eigenvalues and eigenvector.

The main focus of the presentation is on applying PCA to gene expression datasets. It delves into the step-by-step process of performing PCA, starting from data preprocessing, normalization, and scaling, to generating the principal components and analyzing their contributions to the variance. Real-world examples and case studies are presented to illustrate the practical application of PCA in gene expression analysis.

1.1 Literature Survey

Lee et al. [1] propose a machine learning-based approach to integrate diverse sets of big data, including genomic, transcriptomic, proteomic, and clinical data, for the purpose of improving precision medicine for patients with acute myeloid leukemia (AML)

Castillo et al. [2] Proposed pipeline for the correct integration and classification of heterogeneous (Microarray and RNA-seq) biological data.

Nikhil et al. [5] proposed system involves using the genetic score to identify patients with AML and mutated NPM1 who are at high risk of relapse or poor outcomes. This could allow for more personalized treatment approaches.

Mary et al. [6] proposed a system to test for specific genetic mutations or abnormalities that are associated with AML. This testing can help to identify subtypes of AML and guide treatment decisions.

Armstrong et al. [7] proposed system involves analyzing gene expression data from leukemia patients to determine if they have the MLL translocation and to identify the specific gene expression profile associated

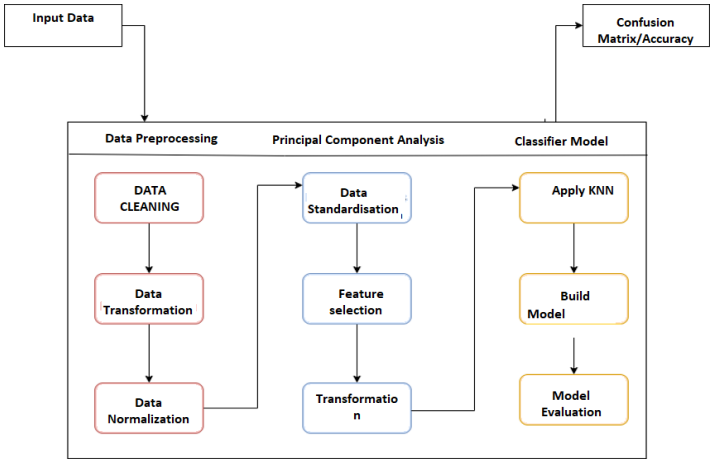
with this genetic abnormality

2. Methodology

Leukemia classification using machine learning algorithms and principal component analysis (PCA) involves a systematic approach to develop a predictive model that can accurately classify leukemia samples based on their gene expression data. The following methodology note outlines the steps involved in this process:

1. Data Collection and Pre-processing: The first step involves collecting leukemia gene expression data from publicly available repositories or through collaborations with other researchers. The data should be pre-processed to remove noise, normalize the data, and transform it to a suitable format for further analysis.
2. Feature Selection: PCA is used to reduce the dimensionality of the data and extract the most relevant features that contribute to the classification of leukemia samples. This involves projecting the high-dimensional gene expression data into a lower-dimensional space, where the principal components (PCs) capture the maximum variance in the data.
3. Model Training and Evaluation: Once the features are selected, various machine learning algorithms such as decision trees, support vector machines, random forests, and neural networks can be trained on the reduced feature set to develop a predictive model. The performance of the model is evaluated using cross-validation and other statistical metrics such as accuracy.
4. Interpretation and Visualization: Finally, the results of the model are interpreted, and the most informative features are visualized to gain insights into the underlying biology of leukemia. This can help identify potential biomarkers and therapeutic targets for the disease.

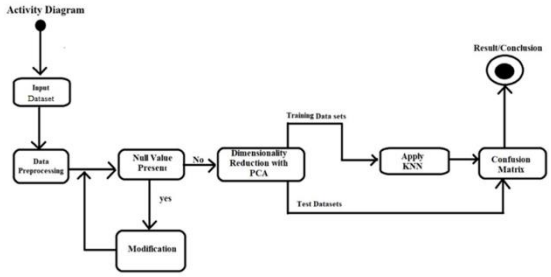
Fig -1: System Design



This is the architecture of our system. It is the actual representation of the algorithm which we implemented. The first step is to read the data set and then it is sent for sampling. Training and testing of the data set is done.

After the feature selection, the data will be sent to the algorithm which is the PCA. The resultant data is stored in test sample data. The prediction of outcome is done based on test sample data the result of the algorithm KNN. It has a feature which validates the results if the result is ALL or AML.

Fig-2- Activity Diagram



In UML, the activity diagram is used to demonstrate the flow of control within the system rather than the implementation. It models the concurrent and sequential activities.

The activity diagram helps in envisioning the workflow from one activity to another. It puts

emphasis on the condition of flow and the order in which it occurs. The flow can be sequential, branched, or concurrent

2.2 Data Structure Design

Pandas DataFrame: The credit card dataset is loaded into a Pandas DataFrame, which is a two-dimensional table-like data structure with labeled rows and columns. It is used to manipulate and analyze the data, as well as perform operations such as selecting and dropping columns, splitting data into input features and output features, and filtering rows.

NumPy arrays: The input features and output feature of the gene expression dataset are converted into NumPy arrays, which are homogeneous arrays of fixed size with efficient element-wise operations. They are used to store and manipulate numerical data and are a popular data structure in machine learning.

StandardScaler: A StandardScaler object is used to scale the input features to have zero mean and unit variance. It is applied to the NumPy arrays of the training and testing sets and is a useful data structure for preprocessing data before applying machine learning models.

2.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a commonly used technique for reducing the dimensionality of large datasets. It helps to transform a large set of variables into a smaller set, while retaining most of the important information present in the larger set. By reducing the number of variables, it becomes easier to analyze and visualize the data, which is particularly useful for machine learning algorithms that require fewer variables to process.

In our study, we used PCA to deal with the problem of having many features or genes in our dataset (over 7,000). This large number of features makes it difficult to train a model that considers all of them. PCA allowed us to pack most of our features into a smaller number of components, while still retaining the most important information. It also helped to discard features that did not show much correlation with the others.

The components generated by PCA may not have a direct representation or meaning, but they are useful for training and visualizing the data. In our case, we compressed the initial 7,129 components into 30 Principal Components using PCA, and then used KNN to train our model based on these 30 components.

Overall, PCA helped us to simplify our dataset and make it more manageable for analysis, without losing too much important information. It is a useful technique for reducing the dimensionality of large datasets and is commonly used in machine learning applications.

2.4 Results

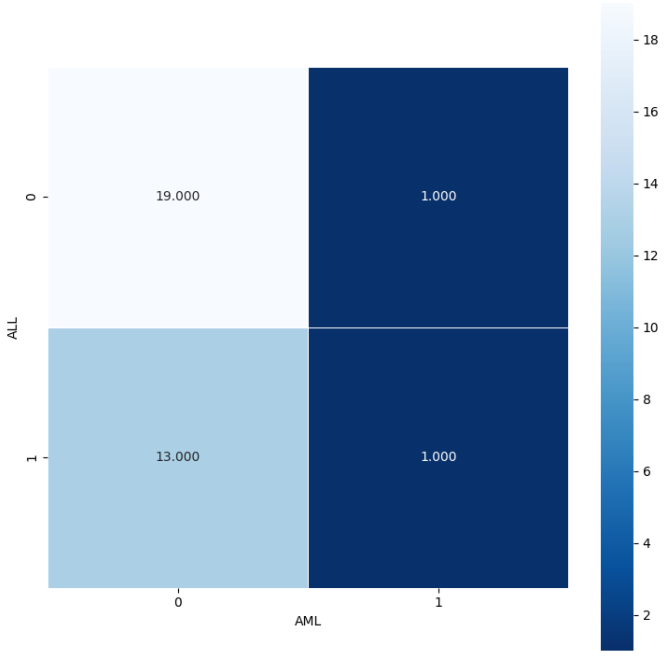
Let's analyze the performance of our model by examining the confusion matrix. The matrix shows the classification results for ALL and AML leukemia classes, represented by the 0 and 1 indices, respectively.

Looking at the results, we see that our model correctly identified 19 out of 20 ALL samples, giving an accuracy rate of 95% for classifying ALL samples. However, the remaining 1 ALL sample was incorrectly classified as AML samples.

In contrast, the performance of the model in classifying AML samples is less satisfactory. Out of 14 AML samples, only one was correctly classified as AML, while the remaining 13 were incorrectly classified as ALL. This indicates that the model has a significant difficulty differentiating between AML and ALL samples, as it tends to misclassify AML samples as ALL samples more frequently than vice versa.

In summary, while our model showed relatively good performance in classifying ALL samples, it needs improvement in differentiating between AML and ALL samples. Further analysis and fine-tuning of the model may be needed to improve its accuracy and reduce the misclassification rate**.**

Fig- Confusion Matrix



3. CONCLUSION

In conclusion, this project has successfully demonstrated the effectiveness and utility of Principal Component Analysis (PCA) in the analysis of gene expression data. Through the application of PCA, we have achieved dimensionality reduction, data visualization, and feature selection, which are crucial steps in gaining insights and understanding complex molecular biology and genetics datasets

3.1 APPLICATION

Disease classification and prognosis: PCA can assist in classifying different disease subtypes based on gene expression patterns. By reducing the dimensionality of gene expression data and visualizing it in a lower-dimensional space, PCA can reveal hidden patterns and relationships that can be used to classify diseases or predict patient outcomes. This application can contribute to personalized medicine and assist in making informed treatment decisions.

3.3 FUTURE ENHANCEMENTS

While PCA is effective in dimensionality reduction, other feature selection methods, such as LASSO, Recursive Feature Elimination, or tree-based methods like Random Forests, could be explored to compare their performance and identify the most relevant features for the given dataset.
Incorporating clinical data, such as patient age, gender, and treatment history, can help to identify factors that influence leukemia prognosis and response to treatment. This can help to develop more accurate predictive models and personalized treatment strategies.

REFERENCES

[1] Lee, S.I., Celik, S., Logsdon, B.A., Lundberg, S.M., Martins, T.J., Oehler, V.G., ... Becker, P.S.: A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nat. Commun.

[2] Castillo D, Galvez JM, Herrera LJ, Rojas F, Valenzuela O, Caba O, Prados J, Rojas I. Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level.

[3] U. K. Dey and M. S. Islam, "Genetic Expression Analysis To Detect Type Of Leukemia Using Machine Learning," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019.

[4]Amirarash Kashef, Toktam Khatibi, Azim Mehrvar,Treatment outcome classification of pediatric Acute Lymphoblastic Leukemia patients with clinical and medical data using machine learning: A case study at MAHAK hospital, Volume 20,2020,100399

[5] Patkar, N., Shaikh, A.F., Kakirde, C. et al. A novel machine-learning-derived genetic score correlates with measurable residual disease and is highly predictive of outcome in acute myeloid leukemia with mutated NPM1. *Blood Cancer J.* **9**, 79 (2019)

[6] Percival ME, Lai C, Estey E, Hourigan CS. Bone marrow evaluation for diagnosis and monitoring of acute myeloid leukemia. Blood Rev.

[7] Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ. MLL

[8] Nazari E, Farzin AH, Aghemiri M, Avan A, Tara M, Tabesh H. Deep Learning for Acute Myeloid Leukemia Diagnosis. J Med Life. 2020 Jul-Sep;13(3):382-387.

[9] Eckardt JN, Bornhäuser M, Wendt K, Middeke JM. Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects. Blood Adv. 2020 Dec 8;4(23):6077-6085

BIOGRAPHIES

Mahalakshmi C.V.
Assistant Professor
Department of C.S.E.
Bangalore Institute of Technology