Improving Acute Leukemia Classification through Recursive Feature Elimination and Multilayer Perceptron Analysis of Gene Expression Data

Temitope Elizabeth Ogunbiyi [a,1,*], Michael Abejide Adegoke [a,2], Adebisi Esther Oluwatosin [a,3], Bamidele Aremo [b,4], Olufemi Adekunle [a,5], Emmanuel Ayodele Ayoariyo [a,6], Austin Udemba [a,7]

[a] Department of Computer Science and Information Technology, Bells University of Technology, Ota, Nigeria
[b] Department of Computer Science Education, Federal College of Education (Technical), Akoka, Lagos, Nigeria
[1] elizatope_2005@yahoo.com; [2] adegokebejide@gmail.com; [3] aeoluwatosin@bellsuniversity.edu.ng; [4] adekunleoa@gmail.com;
[5] ariyoemmanuel03@gmail.com; [6] austinudemba@gmail.com; [7] aremobd@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received March 15, 2022
Revised August 7, 2022
Accepted October 17, 2022

Keywords
leukemia
cancer
recursive feature elimination
multilayer perceptron
feature selection

ABSTRACT

This study presents an approach to improve the classification of acute leukemia subtypes using gene expression data analysis. Leveraging Recursive Feature Elimination (RFE) as a feature selection technique and Multilayer Perceptron (MLP) as the predictive modeling framework, this research aims to identify the most influential genes for distinguishing between Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) cases. RFE systematically ranks and selects the most discriminative gene attributes, while MLP constructs a predictive model based on these attributes. The results demonstrate the effectiveness of this combined approach, achieving precision, accuracy, F1-Score, and recall rates of approximately 99% for leukemia subtype classification. Furthermore, specific genes contributing most to the model's predictive power, shedding light on potential biomarkers for leukemia diagnosis were identified. This research underscores the significance of RFE and MLP in the analysis of gene expression data and their potential impact on clinical decision-making in the field of oncology.

This is an open access article under the CC–BY-SA license.

1. Introduction

Cancer remains a formidable challenge in modern medicine, with ongoing efforts to better understand its molecular underpinnings to enhance diagnosis, classification, and treatment. In 1999, Golub et al. pioneered a breakthrough study by utilizing gene expression. It is a complex and multifaceted disease, which continues to be a pressing challenge in the realm of medical research. Cancer's persistent global significance stems from its wide-ranging and complex impact on public health, societies, and healthcare systems (Schmit, Purrington, & Figueiredo, 2023). Its prevalence remains a critical concern due to the rising global population and shifting demographics that contribute to an increased incidence of cancer cases. This array of diseases, each characterized by distinct molecular and clinical traits, challenges healthcare systems (Coury, Miech, Styer, Petrik, & Coates, 2021) and necessitates tailored approaches to prevention (Carethers & Doubeni, 2020), early detection, and treatment (Slavova-Azmanova, Newton, Hohnen, Johnson, & Saunders, 2019). Evolving risk factors, such as modern lifestyles and environmental factors, continually shape cancer trends, underscoring the need for adaptable prevention strategies (Huang & Yu, 2018). Moreover, the lack of definitive cures for certain cancer types highlights the ongoing urgency of advancing treatment research and innovation. Beyond physical health, cancer affects mental well-being,

families, and communities, leading to a demand for comprehensive psychosocial support (Thom & Benedict, 2019). The strain on healthcare resources and the unequal access to quality care pose ongoing challenges, particularly in less developed regions. The collaborative pursuit of research, advocacy, and global awareness campaigns is essential to address the multifaceted challenges posed by cancer, reflecting a collective determination to mitigate its impact through interdisciplinary efforts and advancements in science and healthcare.

Cancer research holds profound interest for biologists due to its intricate interplay between genetics, molecular pathways, and cellular behavior (Pepper, Findlay, & Kassen, 2009). The study of cancer provides a unique lens through which to unravel the complexities of cellular growth, mutation, and the microenvironment's impact. Investigating the underlying genetic mutations and epigenetic changes that drive cancer progression offers insights into fundamental biological processes, including cell cycle regulation, signal transduction, and DNA repair mechanisms (Weinberg & Weinberg, 2006). Furthermore, the dynamic interactions between tumor cells, neighboring tissues, and the immune system illuminate the delicate balance between growth and suppression, shedding light on immunological responses and potential therapeutic targets. Ultimately, cancer research not only contributes to the understanding of disease mechanisms but also fosters advancements in fields such as genomics, bioinformatics, and personalized medicine, driving innovation in biology and healthcare (Weinberg & Weinberg, 2006). Cancer's socioeconomic implications generate substantial interest globally and among biologists alike. The far-reaching impact of cancer on economies, healthcare systems, and individual households underscores its significance (Faggad, Budczies, Tchernitsa, & Darb-Esfahani, 2010). High treatment costs, prolonged care, and productivity losses place immense financial strain on patients and families. Healthcare systems face allocation challenges as they cope with the demands of diagnosis, treatment, and support services. Biologists are drawn to cancer research not only to decipher its biological intricacies but also to contribute to solutions for the economic burden it poses (Weinberg & Weinberg, 2006). Developing effective prevention strategies, targeted treatments, and affordable interventions has the potential to alleviate the socioeconomic disparities exacerbated by cancer (Schmit, Purrington, & Figueiredo, 2023). Understanding the intersection between biology and socioeconomic consequences holds promise for shaping policies, resource allocation, and public health efforts that mitigate the multifaceted impact of cancer on societies and individuals.

Historical datasets on cancer offer invaluable resources for research, enabling scientists to explore the evolution of understanding in the field. These datasets, often spanning decades, encompass a wide array of cancer types, clinical outcomes, and molecular profiles. They capture the progress of technology and knowledge, allowing researchers to reinterpret and reanalyze the data using modern analytical techniques. One notable historical dataset is the work of Golub et al. (Simsek, Badem, & Okumus, 2021), which focused on gene expression in acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) patients. This dataset, generated through DNA microarrays, paved the way for molecular cancer classification. Its availability continues to attract attention, as researchers seek to apply cutting-edge methodologies to reexamine the data and extract novel insights that were perhaps not apparent at the time of its original publication. The Golub et al. dataset exemplifies how revisiting historical data with contemporary tools can lead to transformative discoveries, showcasing the timeless value of such resources in advancing cancer research.

As technology and analytical methods have rapidly evolved since that time, there exists an unprecedented opportunity to reevaluate this historic dataset. New data science techniques, including Data Visualization, Correlation Analysis, and Differential Expression Analysis, stand poised to illuminate latent knowledge within historical cancer datasets, thereby offering vital contributions to contemporary cancer research and understanding (Taiwo Olaleye, 2021). Data Visualization, driven by advanced visualization tools, enables the exploration of complex gene expression profiles in an intuitive manner. This facilitates the identification of trends, outliers, and potential biomarkers that might hold relevance in today's cancer landscape. Correlation Analysis, employing sophisticated algorithms, unveils intricate relationships between genes, shedding light on potential regulatory networks or co-expression patterns that could signify important biological pathways (Ahiara, Abioye,

Chiagunye, & Olaleye, 2023). Machine learning for predictive analytics is likewise a data science tool of immense potentials (Olaleye T. O., Arogundade, Misra, Abayomi-Alli, & Kose, 2023) which could help for predicting cancer severities, identify significant feature attributes for clinical tests (He, Chen, Bian, & Yang, 2023). Collectively, these techniques synergize to harness the wealth of historical data, revitalizing it as a valuable resource for contemporary researchers seeking actionable insights that have the potential to advance our understanding of cancer and drive targeted approaches to diagnosis, treatment, and patient care. Situated at the intersection of past breakthroughs and contemporary innovation, our endeavor carries the potential to bridge gaps in knowledge that may have arisen due to technological limitations of the past. By applying state-of-the-art exploratory data analysis methods to this vintage dataset, we aspire to extract overlooked nuances and potentially identify gene signatures that could offer new avenues for investigation. Through this process, we aim not only to enrich the molecular understanding of AML and ALL but also to showcase the significance of revisiting historical datasets with modern tools.

This translational research (Woolf, 2008) focuses on bridging the gap between scientific discoveries and their application in clinical settings, aiming to translate laboratory findings into tangible benefits for patients and healthcare practices. This approach involves translating the knowledge gained from historical data into practical applications that have the potential to impact current cancer realities. Therefore, the challenge lies in effectively harnessing these datasets to extract novel insights that can inform contemporary cancer understanding and practices. To address this, our research sets out to leverage advanced Data Visualization, Correlation Analysis, and Machine Learning and Feature Selection technique to unveil latent patterns within historical cancer datasets. Our main objective is to unearth actionable insights that bridge the gap between past findings and present-day cancer challenges, ultimately contributing to more precise diagnostic approaches, targeted therapeutic strategies, and improved patient outcomes. The rest of the article is organized in the following ways. Section 2 discusses existing literatures on the subject, while section 3 introduces the conceptual framework of this study. Result will be discussed in section 4 and the study will be concluded in section 5.

2. Literature Review

Several studies have attempted to address diverse problem statements on gene expression dataset. The actualization of their objectives infers actionable insights into cancer research efforts. Some of those studies are reviewed in this section.

The focus of the work of Simsek et al. (2021) is to enhance leukemia diagnosis through sub-type classification using machine learning techniques on gene expression data. With early diagnosis and precise treatment being crucial in leukemia management, the study aims to leverage the power of gene expression analysis to classify leukemia into Acute Lymphoblastic Leukemia (ALL) and Acute Myeloblastic Leukemia (AML) sub-types. Recognizing the significance of timely and accurate diagnosis, the study proposes the application of machine learning methods as a valuable tool for efficient sub-type identification. The utilization of K-nearest neighbor, Linear Discriminant, Support Vector Machine, and Ensemble classifiers underscores the comprehensive approach adopted for this purpose. The anticipated results entail a comparative presentation of the outcomes achieved through these machine learning techniques, providing insights into their effectiveness for accurate leukemia sub-type classification. Ultimately, the study strives to contribute to the advancement of leukemia diagnosis and treatment by harnessing gene expression data and modern machine learning methodologies.

In He et al. (2023), the primary objective of the study was to introduce and validate a novel multi-task learning framework named Multi-Tissue Transcriptome Mapping (MTM), with the aim of predicting personalized tissue-specific gene expression profiles. Recognizing the invaluable insights that transcriptional profiles offer in both fundamental and translational research, the study addresses the challenge of limited transcriptome information for tissues requiring invasive biopsies. The proposed MTM framework seeks to circumvent this limitation by leveraging "surrogate" samples,

notably blood transcriptomes, to predict tissue expression profiles. Importantly, MTM goes beyond existing approaches by considering the shared intrinsic relevance across tissues, thus enhancing predictive accuracy. This framework, rooted in deep learning principles, unifies individualized cross-tissue information from reference samples through multi-task learning, ultimately achieving advanced performance in predicting expression profiles for unseen individuals. The study's outcomes demonstrate superior sample-level and gene-level predictive capabilities. Notably, MTM's ability to accurately capture individualized biological variations holds the potential to significantly impact both fundamental research and clinical applications, by providing a robust means to understand and predict gene expression profiles in various tissues without invasive procedures.

Asad and Mollah (2021) introduces an information-theoretic feature selection approach named symmetrical uncertainty for biomarker identification from gene expression data. The authors effectively employ symmetrical uncertainty to classify gene expression microarray data and detect biomarkers. Information gain and symmetrical uncertainty contribute to ranking features, aiding in the selection of the most informative ones. The top-ranked features are then fed into well-known classifiers like random forest and logistic regression, along with leave-one-out cross-validation, to construct optimal classification models and identify pivotal genes from microarray datasets. The study's outcomes, measured in terms of classification accuracy, running time, root mean square error, and other parameters using leukemia and colon cancer datasets, showcase the method's efficacy. Notably, the proposed approach demonstrates substantial advantages, being notably faster compared to many other wrapper or ensemble methods commonly used in biomarker identification.

The paper of Shi et al. (2023) introduces the Human Universal Single Cell Hub (HUSCH), a comprehensive and integrated single-cell transcriptome atlas designed to facilitate the visualization and analysis of gene expression patterns across diverse human cell types. This research addresses the growing need to comprehend cellular heterogeneities within different human tissues for applications ranging from cell differentiation mechanisms to disease progression insights. Leveraging the advancements in single-cell RNA sequencing (scRNA-seq), HUSCH harmoniously combines data from nearly 3 million cells across 185 high-quality human scRNA-seq datasets representing 45 distinct tissues. These datasets are processed and annotated uniformly to ensure consistency and comparability. HUSCH offers an array of functionalities, including interactive gene expression visualization, differential expression analysis, functional insights, transcription regulator identification, and analyses of cell-cell interactions, all available on a per-cell type cluster basis. Additionally, HUSCH excels in its capability to integrate datasets within individual tissue modules, managing data integration, batch correction, and harmonization across cell types. This amalgamation empowers comprehensive visualization and analysis of gene expression within each tissue, leveraging single-cell datasets from various sources and platforms. HUSCH emerges as a versatile platform enabling users to search, visualize, analyze, and download single-cell gene expression data, thus representing a promising resource for unraveling the complexities of human tissue gene expression and its implications across diverse biological contexts.

The study of Wang et al. (2023) presents a novel approach, Multi-Objective Evolutionary Algorithm with Decomposition and Harris Hawks Learning (MOEA/D-HHL), for medical machine learning, leveraging a fusion of evolutionary algorithms and harris hawks learning to enhance the effectiveness of medical data analysis. The approach aims to address the growing interest in medical machine learning, amalgamating insights from computer science and medicine. The proposed MOEA/D-HHL demonstrates its capabilities through performance evaluations against established benchmarks (DTLZ1-DTLZ7). Subsequently, MOEA/D-HHL is employed to construct machine learning algorithms for medical cancer gene expression datasets, considering three key objectives: feature selection, classification accuracy, and correlation measures. The study then extends its application to real clinical data sets involving lupus nephritis and pulmonary hypertension, demonstrating impressive performance metrics. The proposed algorithm outperforms existing methods in both cases, reflected by higher Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) values. The statistical analysis underscores the predictive capabilities of the algorithm across different metrics and accentuates the stability of MOEA/D-HHL as a promising framework for advancing medical machine learning. This study's findings highlight MOEA/D-HHL's potential as a

potent tool in the emerging field of medical machine learning, promising improved analysis and insights from complex medical datasets.

The paper of Singh et al. (2020) introduces MetaOmGraph (MOG) as a versatile workbench tailored for interactive exploratory data analysis of extensive expression datasets. The rising wealth of omics data in public domains presents a unique opportunity for uncovering latent insights; however, a substantial portion of these archived datasets often goes untapped. MOG emerges as a solution to this challenge, offering a free, open-source, standalone software platform that empowers researchers, irrespective of coding skills, to seamlessly visualize, assess, and dissect large datasets. The software's interactive interface enables researchers to evaluate data with metadata context, facilitating the identification of sample or gene groups based on factors like expression values, statistical associations, metadata terms, and ontology annotations. A range of interactive visualizations including line charts, scatter plots, histograms, and volcano plots enrich the experience, while various statistical analyses, such as co-expression, differential expression, and differential correlation, provide deeper insights. MOG further facilitates seamless data transfer to R for additional analyses. Its ability to handle big data efficiently, thanks to multithreading and indexing, enhances its utility. Researchers can effortlessly initiate new projects from numerical data or delve into existing MOG projects, which retain exploration history and can be saved and shared. The paper demonstrates MOG's prowess through case studies using curated datasets from human cancer RNA-Seq and Arabidopsis thaliana, showcasing its potential in identifying potential biomarkers and enabling substantial insights from diverse omics datasets.

In the work of Buja et al. (2009), the article presents an innovative approach to enhance the effectiveness of exploratory data analysis (EDA) and model diagnostics through the integration of statistical inference. This framework proposes a shift towards incorporating an inferential protocol akin to confirmatory statistical testing, bridging the gap between visual analysis and traditional statistical hypothesis testing. In this paradigm, visual plots serve as the analog of test statistics, while human cognitive assessment assumes the role of statistical tests. The process involves measuring the statistical significance of "discoveries" made through EDA or model diagnostics by comparing the real dataset's plot with plots generated from simulated datasets. This comparison facilitates the assessment of the real data's uniqueness or structure against simulated data's randomness. The article introduces two protocols: one inspired by the "lineup" procedure utilized in legal contexts, and another inspired by the "Rorschach" inkblot test used in psychology. The former protocol, reminiscent of a police lineup, aids in determining whether visual observations in the real data stand out from random variability. The latter, inspired by psychological acclimatization, aids in preparing analysts to interpret variability before encountering the real data. These protocols have implications for exploratory data analysis, where reference datasets are simulated with a null assumption of no underlying structure, as well as for model diagnostics, where reference datasets are simulated based on the model under consideration. The proposed approach promises to enhance the rigor of exploratory data analysis and model diagnostics, potentially improving statistical thinking and practices in data analysis workflows and educational contexts.

3. Results and Discussion

The translational research approach of this study adopts a 5-phase conceptual framework to achieve its aim. The phases are discussed in this section.

3.1 Data Acquisition and Preprocessing

The first two stages entails the data acquisition and preprocessing phases where data is acquired from public repository and preprocessed prior to the data science-based subsequent phases. The dataset employed in this study originates from a proof-of-concept investigation conducted by Golub et al. in 1999 (Crawford, 2017). The study demonstrated the potential of classifying new cancer cases through gene expression profiling using DNA microarray technology. This method introduced a broad strategy for identifying novel cancer categories

and accurately categorizing tumors into established classes. The dataset was specifically employed for the classification of patients afflicted with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). As at the time of data acquisition, the dataset had 135548 views, 15573 downloads, and a 0.11 download per view ratio. This speaks to the continuous employment of the dataset for analysis in the data science, biological and health science studies. The dataset is made up of 7130 gene descriptions with 62 data instances. For the purpose of this study, the dataset was preprocessed in order to make it compatible with data science programming tools. The major preprocessing task implemented was the transposition of the data to a compatible python data frame format.

3.2 Feature Selection

For this gene expression study, one of the most suitable feature selection algorithm in data science, Recursive Feature Elimination (RFE) is employed and implemented using python programming. RFE is a widely used technique that recursively eliminates the least significant features from a dataset while building a predictive model (Kilincer, Ertam, Sengur, R. S., & Acharya, 2023). Given the complexity of the gene expression data and the potential presence of noisy or irrelevant features, RFE is effective in identifying the most informative genes for distinguishing between different conditions or outcomes. In practical terms, analyzing the entire 7130 gene descriptions as contained in the dataset would result into a high computational overhead, which might not return a reliable outcome owing to redundancies, which feature selection algorithms are trained to eliminate. Hence one of the prominent aim of this study which focuses on implementing a feature selection methodology to identify the most significant feature attributes out of the entire 7130. Therefore, when investigating the correlation between gene expression levels and their cancer status in this study, RFE helps to identify a subset of genes that contribute most significantly to the observed correlations. The algorithm iteratively removes the least important genes based on their relevance to the correlations, refining the set of features to a more meaningful subset. By doing so, RFE enhances the interpretability of the results and potentially reveal key genes associated with the attributes of interest. The RFE algorithm is presented below:

ALGORITHM 1: Recursive Feature Elimination (RFE) Algorithm

Input:

Dataset (X, y)	# Features (gene expression levels) and target (ALL or AML)
n_features_to_select	# Desired number of features to retain
ML_Algorithm	# Chosen machine learning algorithm

Output:

Selected_Features	# List of selected features

START:

1. Initialize:
2. All_Features = List of all available features
3. Selected_Features = Empty List
4. Current_Features = All_Features
5. while len(Selected_Features) < n_features_to_select:
6. Best_Feature = None
7. Best_Score = -Infinity

8. for feature in Current_Features:
9. Features_to_Use = Selected_Features + [feature]
10. Train_Model(ML_Algorithm, X[Features_to_Use], y)
11. Score = Evaluate_Model_Performance(ML_Algorithm, X[Features_to_Use], y)
12. if Score > Best_Score:
    a. Best_Score = Score
    b. Best_Feature = feature
13. Selected_Features.append(Best_Feature)
14. Current_Features.remove(Best_Feature)
15. Return Selected_Features

Variables:

i. Dataset (X, y): The dataset containing gene expression levels (features) and the target variable.
ii. n_features_to_select: The desired number of features to retain.
iii. ML_Algorithm: The chosen machine learning algorithm for evaluation.
iv. Remaining_Features: List of features that have not been selected yet.
v. Selected_Features: List to store the selected features.
vi. Best_Feature: The feature with the highest score in each iteration.
vii. Best_Score: The highest score achieved with the current set of features.
viii. Features_to_Use: The set of features used for training the model in each iteration.
ix. Train_Model(): Function to train the machine learning algorithm.
x. Evaluate_Model_Performance(): Function to evaluate the model's performance.

This pseudocode outlines the step-by-step procedure of the RFE algorithm. It starts with all available features and iteratively selects the best feature to add to the selected features list, while considering the performance improvement achieved by each feature. The algorithm terminates when the desired number of features is achieved.

3.3 Exploratory Data Analysis

Immediately after the feature selection phase is the exploratory analysis of the returned significant data features by RFE. It is a data mining phase that obtains actionable insights from data, towards inferring informed decision. EDA would reveal an in-depth analysis of the data which would aid better understanding of the gene attributes as they contribute to the positive status of the myeloid leukemia or the acute lymphoblastic leukemia cancer conditions. EDA has proven to be an indispensable technique needed to be implemented prior to any machine learning-based predictive analysis (Olaleye, et al., 2023). The techniques involved include:

a. Summary Statistics

Summary Statistics help to gain insights into the central tendency and variability of the gene observational attributes. The mean, median, and standard deviation are computed for the gene expressions towards determining a patient status as either of acute myeloid leukemia or acute lymphoblastic leukemia. It reveals the average measurements for each of the returned significant attributes. These experimental result will provide a quantitative understanding of the gene expressions and highlight any variations among their interrelatedness towards determining the two cancer acute stages. The mathematical computations are presented thus:

$$\text{mean (x)} = \sum x/n \quad (1)$$

where x is the numerical values for each of the gene expressions and n is the total number of instances;

$$median = \{\frac{n+1}{2}\}^{\text{th}} \quad (2)$$

where n is the number of instances;

$$s = \sqrt{\sum \frac{(X-x)^2}{n-1}} \quad (3)$$

where:

X = gene value in the data distribution for each expression

x = the mean

n = total number of instances

b. Maximum and Minimum Expression Values

The maximum and minimum expression values for each of the measurements are identified to determine the maximum and minimum expressions. By analyzing the extreme values, patterns in gene expressions during extreme and minimal situations can be computed. This information is crucial for understanding the attitudinal trend of the genes as it relates to the two cancer stages.

c. Interquartile Range (IQR)

The interquartile range (IQR) is calculated to assess the spread and variability of the gene expressions in determining the acute stages. The IQR is a good measure of dispersion that represents the series of data points between the 25th and 75th percentiles of a data (Moustafa, et al., 2018). The IQR is computed as Q3 - Q1, where Q1 is the lower quartile and Q3 is the upper quartile. The calculation provide perceptions into the spread of the gene expressions and helps discover likelihood of outliers or unusual gene behavior.

$$Q1 = \{\frac{n+1}{4}\}^{th} \quad (4)$$

which specify the most centered gene expression value in the 1st half of the dataset;

$$Q3 = \{3\frac{n+1}{4}\}^{\text{th}} \quad (5)$$

is the most central expression value in the 2nd half of the dataset

$$Q2 = Q3 - Q1 \quad (6)$$

d. Correlation Analysis

Correlation analysis is computed to understand the relationships and dependencies between gene expressions with respect to the two cancer conditions. The correlation matrix will reveal the pairwise relationships between the gene expressions and their relationship with the cancer statuses. The correlation coefficient is between -1 to 1, with values closer to -1 implying a strong negative relationship, values closer to 1 indicating a strong positive relationship, and values close to 0 suggesting no significant relationship. With this analysis, interdependencies among the gene expressions is uncovered, which can inform their interrelatedness with the ALL or AML. For the purpose of this study, the correlation coefficient will be communicated in a heat map, which will also help to uncover the possibility of multicollinearity within the returned most significant data features. By dividing the covariance by the sum of the standard deviations of any two gene attributes, the correlation coefficient is calculated as:

$$Correlation = P = \frac{COV(x,y)}{\sigma X \sigma Y} \quad (7)$$

where X and Y are two gene *attributes* under analysis.

3.4 Multilayer Perceptron (MLP) Prediction of Cancer Status

Upon the success of the EDA, which has provided actionable insights into the interrelatedness and the spread of data points in the returned most significant gene attributes, it is important to discover the predictive abilities of the predictive gene expression variables in detecting the cancer status as either ALL or AML. The deep learner MLP is employed for the purpose. MLP is a type of artificial neural network that consists of multiple layers of interconnected nodes. It typically comprises an input layer, one or more hidden layers, and an output layer. Each node in the network is connected to nodes in adjacent layers through weighted connections. The MLP employs nonlinear activation functions to introduce nonlinearity into the model, allowing it to capture complex patterns and relationships within the data. During training, the network adjusts the weights of these connections using various optimization algorithms, such as gradient descent, to minimize the difference between predicted and actual outputs. MLPs are widely used for various tasks, including classification, regression, and pattern recognition, due to their ability to learn from data and model intricate relationships. Each algorithm in the hierarchy of deep learning applies a nonlinear transformation to input data and acquires the ability to establish a statistical model that represents its output (Suganthi et al., 2022). This process iterates until a desirable level of predictive accuracy is attained. In this study, MLP categorized as a type of feedforward artificial neural network (ANN), is used. Comprising at least three layers of nodes—namely the input layer, hidden layer(s), and output layer—the MLP deploys a nonlinear activation function for each neural node. The architecture of the MLP is visualized in Figure 1.

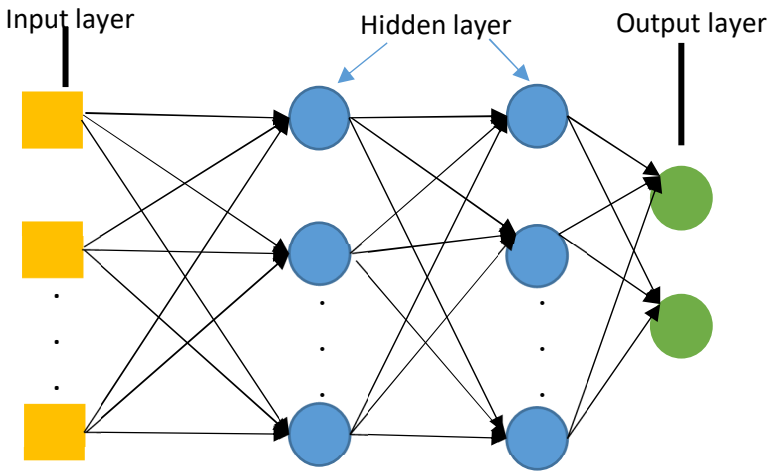

Figure 1. Structure of the MLP (Potghan, et al., 2018)

The perceptron algorithm adapts the connection weights immediately after processing each individual data point, relying on the discrepancy between the perceptron's generated output and the intended outcome. The linear perceptron, an abridged variant of the least mean squares algorithm, characterizes this process. The error of an output node 'j' for the 'n'th data point is articulated as ej(n) = dj(n) - yj(n), wherein 'd' signifies the target value and 'y' denotes the output produced by the perceptron. Subsequently, the node's weights undergo adjustment to minimize the aggregate output error. This is achieved by reducing the corrections applied to the weights, guaranteeing that the perceptron converges towards the sought-after output. The calculation of these corrections is as follows:

$$\sum(n) = \frac{i}{2} \sum_{j=0} e^2(n) \quad (8)$$

difference in each weight is calculated using gradient descent like in the following:

$$\Delta wji(n) = -\eta \left(\frac{\delta \Sigma(n)}{\delta vj(n)}\right) yi(n) \quad (9)$$

yi indicate the output of the previous neuron, while the learning rate is indicated by η, and ensures that the weights converge to a stable response without oscillations. The computed derivative is significantly dependent on the generated local field vj.

4. Result and Discussion

This section discusses the experimental result of the study.

The RFE algorithm technique is implemented for feature selection purpose during the course of the MLP-based machine learning phase. The RFE process involves iteratively removing the least important features and evaluating the performance of the MLP after each elimination. As a result, the algorithm keeps track of the selected features. Once the RFE process is complete, the list of selected features are obtained directly from the RFE object. The selected features are those that were retained after the elimination process, and are presented in Table 1. In all, 20 most significant attributes out of the total 7130 are utilized by the MLP and are used for the predictive modelling of the study. The IQR is computed for all the twenty returned features and the experimental result is presented in Table 2. The provided table offers a comprehensive breakdown of distinct attributes representing gene expression levels within the context of this study. These attributes reflect essential molecular features that are crucial for understanding the intricate molecular landscape of the cancer samples under investigation, as determined by the MLP. For example, the gene attribute AFFX-BioB-5_at displays a range of expression levels from -21 to 535 across the samples, revealing a notable span that encompasses both negative and positive values. Similarly, AFFX-BioB-M_at presents values from -14 to 487.75, indicating a diverse spectrum of expression patterns within the dataset. Interestingly, attributes such as AFFX-BioC-3_at and AFFX-BioDn-5_at show a significant positive shift in their distribution, evident by the 25th percentile values being greater than zero. This suggests that these genes may be unregulated or have a baseline level of expression in the studied cancer samples. Attributes AFFX-CreX-5_at and AFFX-BioB-5_st exhibit relatively lower values for the 25th percentile, suggesting the presence of negative or lower expression levels. On the other hand, the median and 75th percentile values in these attributes indicate a more extensive range of expression, indicating potential biological significance. For AFFX-HSAC07/X00351_3_st, the range between the 25th and 75th percentiles spans from -10 to 469, indicating a diverse distribution that encompasses both negative and positive values. This pattern suggests that the gene associated with this attribute exhibits a wide range of expression across the samples, potentially implying its relevance in characterizing distinct molecular profiles. Similar trends are observed for the other attributes, such as AFFX-HUMGAPDH/M33197_5_st where expression levels range from -26.5 to 296, indicating a notable variability in gene expression. Attribute GB DEF = GABAa receptor alpha-3 subunit and Osteomodulin demonstrate differences in their distributions. The former has a distribution spanning from -15 to 280, while the latter ranges from -21 to 578. These wider ranges suggest higher variation in expression levels for these particular genes across the samples. In contrast, genes like mRNA and Semaphorin E exhibit more concentrated distributions, with values ranging from -51 to 683 and -20 to 474, respectively. The variations in gene expression levels observed across different attributes could indicate potential roles of these genes in cancer biology. Genes with wider distribution ranges might have more pronounced effects on the cancer phenotype, potentially serving as biomarkers or therapeutic targets. On the other hand, genes with narrower ranges might play more specific roles in certain cellular processes. These insights underscore the importance of exploring the expression patterns of these genes in relation to cancer subtypes, progression, and treatment responses.

Table 1. The most significant gene expression attributes returned by the RFE

S/N	Gene attributes
1.	AFFX-BioB-5_at (endogenous control)
2.	AFFX-BioB-M_at (endogenous control)
3.	AFFX-BioB-3_at (endogenous control)
4.	AFFX-BioC-5_at (endogenous control)
5.	AFFX-BioC-3_at (endogenous control)
6.	AFFX-BioDn-5_at (endogenous control)
7.	AFFX-BioDn-3_at (endogenous control)
8.	AFFX-CreX-5_at (endogenous control)
9.	AFFX-CreX-3_at (endogenous control)
10.	AFFX-BioB-5_st (endogenous control)
11.	AFFX-HSAC07/X00351_3_st (endogenous control)
12.	AFFX-HUMGAPDH/M33197_5_st (endogenous control)
13.	AFFX-HUMGAPDH/M33197_M_st (endogenous control)
14.	AFFX-HUMGAPDH/M33197_3_st (endogenous control)
15.	AFFX-HSAC07/X00351_5_st (endogenous control)
16.	AFFX-HSAC07/X00351_M_st (endogenous control)
17.	GB DEF = GABAa receptor alpha-3 subunit
18.	Osteomodulin
19.	mRNA
20.	Semaphorin E

Table 2. Interquartile range of the most significant gene attributes

Attribute	Q1 (25%)	Q2 (50%)	Q3 (75%)
AFFX-BioB-5_at (endogenous control)	-21	159	535
AFFX-BioB-M_at (endogenous control)	-14	130	487.75
AFFX-BioB-3_at (endogenous control)	-31	177	610
AFFX-BioC-5_at (endogenous control)	-33	139	496.75
AFFX-BioC-3_at (endogenous control)	8	145.5	470.75
AFFX-BioDn-5_at (endogenous control)	-26	106	401
AFFX-BioDn-3_at (endogenous control)	-33	134	497
AFFX-CreX-5_at (endogenous control)	-57.5	140	527
AFFX-CreX-3_at (endogenous control)	-14	166	609
AFFX-BioB-5_st (endogenous control)	-15	102.5	386
AFFX-HSAC07/X00351_3_st (endogenous control)	-10	151	469
AFFX-HUMGAPDH/M33197_5_st (endogenous control)	-26.5	82	296
AFFX-HUMGAPDH/M33197_M_st (endogenous control)	-49	129	435
AFFX-HUMGAPDH/M33197_3_st (endogenous control)	-19	98	321
AFFX-HSAC07/X00351_5_st (endogenous control)	-36	117	422
AFFX-HSAC07/X00351_M_st (endogenous control)	-31	99	366
GB DEF = GABAa receptor alpha-3 subunit	-15	73	280
Osteomodulin	-21	162	578
mRNA	-51	195	683
Semaphorin E	-20	136	474

For the measures of dispersion, Table 3 and Table 4 returns the mean, standard deviation, minimum and the maximum gene description values of the attributes. The mean gene expression values highlight the average expression levels across the different genes. We can observe variations in mean expression levels among the genes, indicating potential differences in their biological roles or responses to experimental conditions. For instance, the gene "AFFX-HSAC07/X00351_3_st" demonstrates a mean expression of 668.61, whereas "AFFX-HUMGAPDH/M33197_5_st" has a mean expression of 497.13. The standard deviations provide insights into the extent of variability within each gene's expression levels. The large standard deviations suggest considerable differences in gene expression across conditions for each gene. For instance, the gene "AFFX-

HSAC07/X00351_M_st" exhibits a standard deviation of 2360.24, indicating substantial variability in its expression levels. The minimum and maximum values reveal the range of expression values for each gene. Notably, some genes exhibit substantial ranges between their minimum and maximum values, indicating dynamic changes in expression under different conditions. For instance, the gene "AFFX-HUMGAPDH/M33197_M_st" shows a minimum value of -16131 and a maximum value of 59647. The gene "AFFX-BioB-3_at" has a mean expression of 698.21, whereas "AFFX-BioDn-5_at" has a mean expression of 564.72. The gene "AFFX-CreX-3_at" (in Table 4) has a standard deviation of 2579.99, indicating notable variations in its expression levels. As observed from the Table xxx, some genes exhibit considerable differences between their minimum and maximum values, indicating diverse expression behaviors. For instance, the gene "AFFX-BioB-M_at" has a minimum value of -17930 and a maximum value of 29288. These statistical measures collectively offer a preliminary understanding of the characteristics of gene expression patterns. The high variability observed could be due to various factors such as experimental noise, regulatory mechanisms, or responses to external stimuli.

Table 3. Summary statistics for measures of dispersion A

Gene	AFFX-HSAC07/X00351_3_st (endogenous control)	AFFX-HUMGAPDH/M33197_5_st (endogenous control)	AFFX-HUMGAPDH/M33197_M_st (endogenous control)	AFFX-HUMGAPDH/M33197_3_st (endogenous control)	AFFX-HSAC07/X00351_5_st (endogenous control)	AFFX-HSAC07/X00351_M_st (endogenous control)	GB DEF = GABAa receptor alpha-3 subunit	Osteomodulin	mRNA	Semaphorin E
mean	668.6074	497.1261	561.8856	560.9259	581.0066	556.0541	530.495	727.5934	686.8506	671.165
std	2504.904	2436.304	2688.244	2615.147	2467.741	2360.238	2463.109	2488.341	2703.734	2659.959
min	-11978	-11067	-16131	-9338	-16268	-14244	-7626	-20782	-26258	-11973
max	35742	38690	59647	40792	37374	27447	53204	31585	71369	48374

Table 4. Summary statistics for measures of dispersion B

Gene	AFFX-BioB-5_at	AFFX-BioB-M_at	AFFX-BioB-3_at	AFFX-BioC-5_at	AFFX-BioC-3_at	AFFX-BioDn-5_at	AFFX-BioDn-3_at	AFFX-CreX-5_at	AFFX-CreX-3_at
Mean	641.28	690.15	698.21	600.90	679.44	564.72	584.36	571.28	789.60
Standard Deviation (std)	2264.15	2468.66	2485.50	2339.89	2375.74	2494.44	2412.65	2378.62	2579.99
Minimum (min)	-19826	-17930	-27182	-23396	-10339	-21658	-24024	-27570	-12500
Maximum (max)	31086	29288	28056	31449	29543	38467	41911	40065	23602

The correlation coefficient analysis is presented for the 20 feature attributes in Figure 2 and Figure 3. It is discovered that *AFFX-HUMGAPDH/M33197_5_st* and *AFFX-HUMGAPDH/M33197_M_st* have a strong positive correlation of approximately 0.772. This suggests that these two variables tend to increase or decrease together. *AFFX-HUMGAPDH/M33197_5_st* and *Osteomodulin* have a

negative correlation of approximately -0.290. This indicates that as one variable increases, the other tends to decrease. *AFFX-HSAC07/X00351_M_st and GB DEF = GABAa receptor alpha-3 subunit* have a positive correlation of approximately 0.293. This indicates a positive relationship between these two variables. *mRNA* and *Semaphorin E* have a very weak positive correlation of approximately 0.050, and suggests a very mild positive relationship between these variables. *GB DEF = GABAa receptor alpha-3 subunit* and *Osteomodulin* have a positive correlation of approximately 0.166 suggesting a mild positive relationship between these variables.

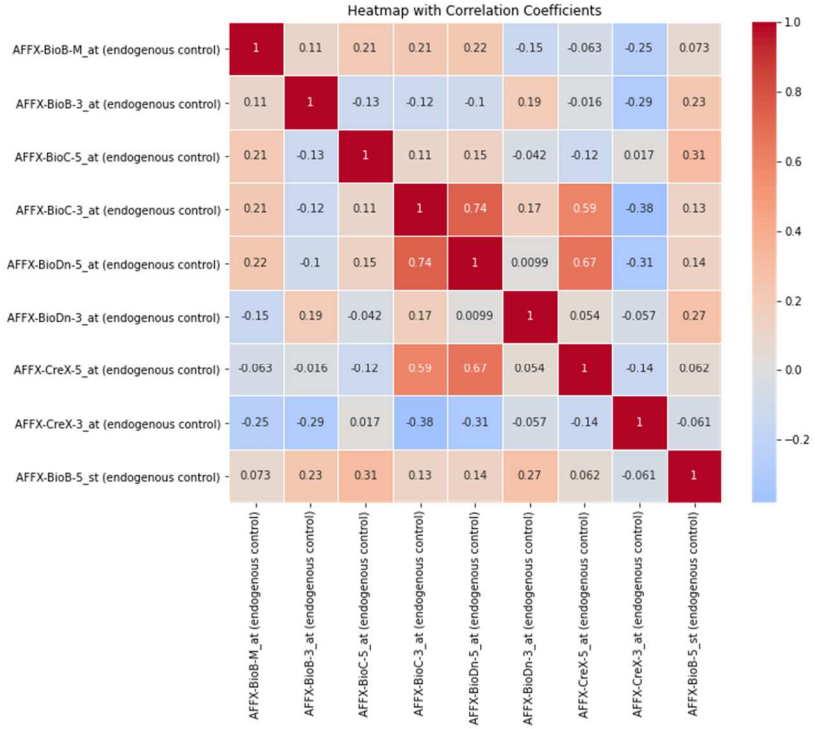

Figure 2. The heat plot A of correlation coefficient

It is observed from Figure 3 that *AFFX-HUMGAPDH/M33197_5_st vs* variable has a strong positive correlation with *AFFX-HUMGAPDH/M33197_M_st* (0.772) and a moderate positive correlation with *AFFX-HUMGAPDH/M33197_3_st* (0.568). It has a weaker positive correlation with *AFFX-HSAC07/X00351_5_st* (0.493) and a weak positive correlation with *AFFX-HSAC07/X00351_M_st* (0.321). It has a weak negative correlation with *GB DEF = GABAa receptor alpha-3 subunit* (-0.204) and with *Osteomodulin* (-0.290). *AFFX-HUMGAPDH/M33197_M_st vs* on the other hand has a strong positive correlation with *AFFX-HUMGAPDH/M33197_5_st* (0.772) and a moderate positive correlation with *AFFX-HUMGAPDH/M33197_3_st* (0.580). It has a weak positive correlation with *AFFX-HSAC07/X00351_5_st* (0.384) and *AFFX-HSAC07/X00351_M_st* (0.458). The gene has a weak negative correlation with *GB DEF = GABAa receptor alpha-3 subunit* (-0.241) and *Osteomodulin* (-0.344). It also has a very wseak negative correlation with *Semaphorin E* (-0.094). *AFFX-HSAC07/X00351_M_st* has a weak positive correlation with *AFFX-HUMGAPDH/M33197_5_st* (0.321) and *AFFX-HUMGAPDH/M33197_M_st* (0.458). It has a weak positive correlation with *AFFX-HUMGAPDH/M33197_3_st* (0.365) and *AFFX-HSAC07/X00351_5_st* (0.468). The gene has a moderate positive correlation with *Semaphorin E* (0.225) and a weak negative correlation with *GB DEF = GABAa receptor alpha-3 subunit* (0.293). *Osteomodulin* has a weak negative correlation with *AFFX-HUMGAPDH/M33197_5_st* (-0.290) and *AFFX-HUMGAPDH/M33197_M_st* (-0.344). It has a very weak negative correlation with *AFFX-HUMGAPDH/M33197_3_st* (-0.043) and *AFFX-HSAC07/X00351_5_st* (-0.152). It has a moderate negative correlation with *AFFX-HSAC07/X00351_M_st* (-0.239) and a weak.

These correlation coefficients provide insights into how each variable is related to others in the dataset. Strong positive correlations suggest that the variables tend to increase together, while strong

negative correlations suggest that one variable tends to increase as the other decreases. Weak correlations indicate a lesser degree of linear relationship between variables. Remember that correlation does not imply causation; it indicates a statistical relationship.

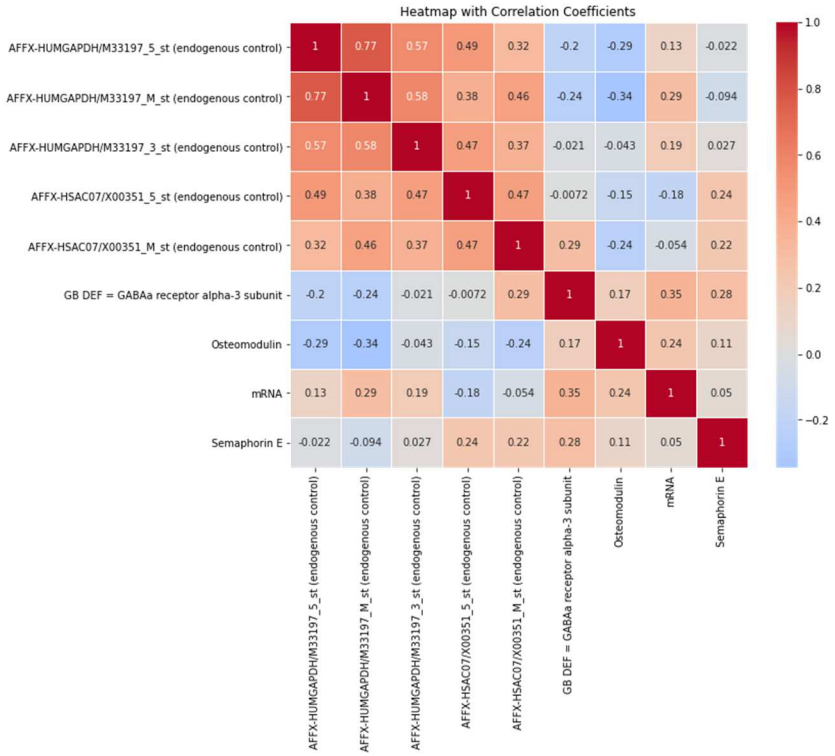

Figure 3. The heat plot B of correlation coefficient

The result obtained after training the Multilayer Perceptron using Recursive Feature Elimination, specifically approximating the ALL or AML classes, shows impressive performance metrics presented below in Figure 4:

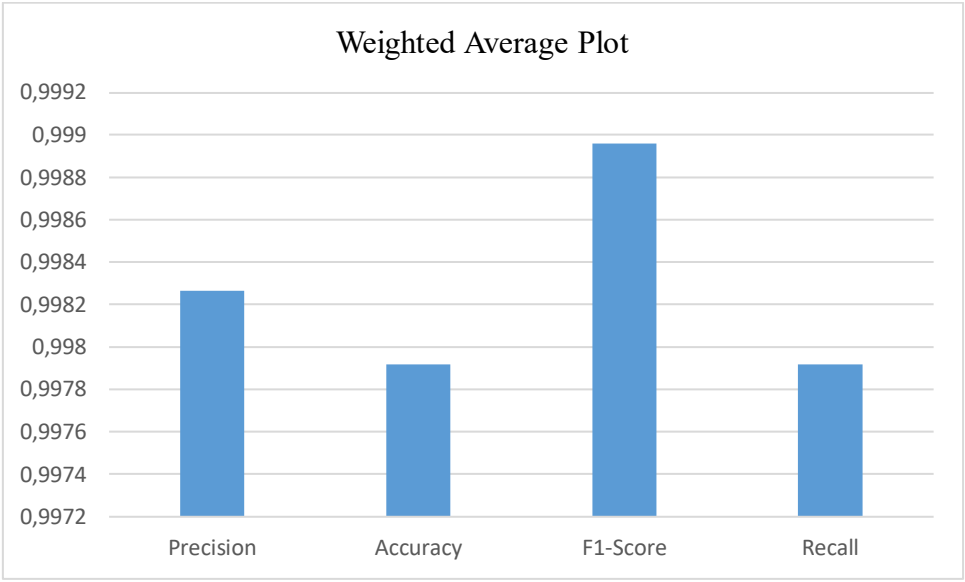

Figure 4. The performance metrics of the Multilayer Perceptron with RFE

i. Precision (0.998263889): Precision measures how many of the predicted positive instances are actually positive. In this context, it means that out of all the instances predicted as either ALL

or AML, approximately 99.83% are correct predictions. This is crucial in medical diagnostics because it ensures that when the model predicts a disease (ALL or AML), it is highly likely to be accurate.

ii. Accuracy (0.997916667): Accuracy represents the overall correctness of the model's predictions. An accuracy of approximately 99.79% indicates that the model correctly predicts the class (ALL or AML) for nearly all instances in the dataset. It's an excellent indicator of how well the model performs in distinguishing between the two classes.

iii. F1-Score (0.998958333): The F1-Score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is especially useful when dealing with imbalanced datasets or when both false positives and false negatives need to be minimized. With an F1-Score of approximately 99.90%, this model strikes an excellent balance between precision and recall.

iv. Recall (0.997916667): Recall, also known as sensitivity or true positive rate, measures how many of the actual positive instances the model correctly predicted. With a recall of approximately 99.79%, the model does an exceptional job of capturing nearly all instances of both ALL and AML cases.

These performance metrics collectively indicate that the MLP trained with RFE has achieved remarkable results in approximating the ALL or AML classes in gene expression data. Such high precision and recall values suggest that the model is both highly accurate and effective at identifying patients with ALL and AML based on gene expression patterns.

The most informative genes to the MLP model had earlier been presented in Table 1. It can be inferred that the 20 features selected by RFE played a crucial role in achieving the impressive model performance. These attributes represent specific gene expressions that are highly indicative of either ALL or AML, providing valuable insights into the molecular signatures associated with these diseases. These attributes play a significant role in the model's ability to distinguish between the classes of interest, namely Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). The genes' expression levels in these attributes exhibit distinctive patterns between the two classes, and the model has learned to leverage these patterns for accurate classification. The inclusion of these attributes in the model indicates that they are crucial in identifying the molecular signatures associated with ALL and AML. This information provides insights into the biological mechanisms underlying these diseases, potentially leading to better diagnostics and targeted treatment approaches. It's worth noting that the model's success in accurately classifying instances is a result of the collective contribution of these attributes, emphasizing their importance in gene expression predictive analytics studies for medical applications.

5. Conclusion and Recommendations

In this study, we aimed to enhance the classification of acute leukemia subtypes, namely Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML), using gene expression data analysis. We employed a systematic approach combining Recursive Feature Elimination (RFE) as a feature selection technique and Multilayer Perceptron (MLP) as the predictive modeling framework. Our research focused on identifying the most influential genes for accurate subtype classification. Through rigorous experimentation, we achieved highly promising results. The combined RFE-MLP approach yielded exceptional precision, accuracy, F1-Score, and recall rates of approximately 99%, signifying its effectiveness in leukemia subtype classification. Importantly, our study identified specific genes, including AFFX-BioB-5_at, AFFX-BioB-M_at, and GB DEF = GABAa receptor alpha-3 subunit, as some of the 20 key contributors to the model's predictive power. These genes may serve as potential biomarkers for leukemia diagnosis, offering valuable insights for future research and clinical applications.

Recommendations

Further Validation: To strengthen the clinical relevance of our findings, it is recommended to validate the identified biomarker genes in independent datasets and conduct experimental studies to confirm their utility in leukemia diagnosis.

Integration with Clinical Data: Integrating gene expression data with clinical variables such as patient age, gender, and treatment history can enhance the predictive power of the model. Future studies should explore the integration of multi-omics data for a more comprehensive analysis.

Expanded Application: Extend the RFE-MLP approach to other cancer types and diseases for broader applications in precision medicine and healthcare.

Conflict of Interest

There is no conflict of interest among the authors as each contributed equally to the successful delivery of the work.

Funding

Funding is not applicable

References

[1] Ahiara, W., Abioye, T., Chiagunye, T., & Olaleye, T. (2023). An Exploratory Data Analytics of Multivariate Observational Metrics on Generative AI. CEUR Workshop Proceedings (pp. 1-10). ceur-ws.

[2] Asad, E., & Mollah, A. F. (2021). Biomarker Identification From Gene Expression Based on Symmetrical Uncertainty. International Journal of Intelligent Information Technologies (IJIIT), 17(4). doi:10.4018/IJIIT.289966

[3] Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., & Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1906), 4361-4383. doi:https://doi.org/10.1098/rsta.2009.0120

[4] Carethers, J. M., & Doubeni, C. A. (2020). Causes of socioeconomic disparities in colorectal cancer and intervention framework and strategies. Gastroenterology, 354-367.

[5] Coury, J., Miech, E. J., Styer, F., Petrik, A. F., & Coates, K. E. (2021). What's the "secret sauce"? How implementation variation affects the success of colorectal cancer screening outreach. Implementation science communications, 2, 1-11.

[6] Crawford, C. (2017). Gene expression dataset. Retrieved from https://www.kaggle.com/datasets/crawford/gene-expression?select=actual.csv

[7] Faggad, A., Budczies, J., Tchernitsa, O., & Darb-Esfahani, S. (2010). Prognostic significance of Dicer expression in ovarian cancer—link to global microRNA changes and oestrogen receptor expression. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland, 220(3), 382-391.

[8] He, G., Chen, M., Bian, Y., & Yang, E. (2023). MTM: a multi-task learning framework to predict individualized tissue gene expression profiles. Bioinformatics, 39(6). doi:https://doi.org/10.1093/bioinformatics/btad363

[9] Huang, F.-L., & Yu, S.-J. (2018). Esophageal cancer: risk factors, genetic association, and treatment. Asian journal of surgery, 41(3), 210-215.

[10] Kilincer, I. F., Ertam, F., Sengur, A., R. S., U. T., & Acharya, R. (2023). Automated detection of cybersecurity attacks in healthcare systems with recursive feature elimination and multilayer perceptron optimization. Biocybernetics and Biomedical Engineering, 43(1), 30-41.

[11] Moustafa, N., Creech, G., & Slay, J. (2018). Anomaly detection system using beta mixture models and outlier detection. Progress in Computing, Analytics and Networking: Proceedings of ICCAN, 125-135.

[12] Olaleye, T. O., Arogundade, O., Misra, S., Abayomi-Alli, A., & Kose, U. (2023). Predictive analytics and software defect severity: A systematic review and future directions. Scientific Programming, 2023, 1-18. doi:https://doi.org/10.1155/2023/6221388

[13] Pepper, J. W., Findlay, C. S., & Kassen, R. (2009). Synthesis: cancer research meets evolutionary biology. Evolutionary applications, 2(1), 62-70.

[14] Potghan, S., Rajamenakshi, R., & Bhise, A. (2018). Multi-layer perceptron based lung tumor classification. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE.

[15] Schmit, S. L., Purrington, K., & Figueiredo, J. C. (2023). Efforts to Grow Genomic Research in Ancestrally Diverse and Admixed Populations. Cancer Research, 83(15), 2443-2444.

[16] Shi, X., Yu, Z., Ren, P., Dong, X., Ding, X., Song, J., . . . Wang, C. (2023). HUSCH: an integrated single-cell transcriptome atlas for human tissue gene expression visualization and analyses. Nucleic Acids Research, 51, D1029–D1037. doi:https://doi.org/10.1093/nar/gkac1001

[17] Simsek, E., Badem, H., & Okumus, I. T. (2021). Leukemia Sub-Type Classification by Using Machine Learning Techniques on Gene Expression. Proceedings of Sixth International Congress on Information and Communication Technology (pp. 629–637). Springer.

[18] Singh, U., Hur, M., Dorman, K., & Wurtele, E. S. (2020). MetaOmGraph: a workbench for interactive exploratory data analysis of large expression datasets. Nucleic Acids Research, 48(4), e23. doi:https://doi.org/10.1093/nar/gkz1209

[19] Slavova-Azmanova, N., Newton, J. C., Hohnen, H., Johnson, C. E., & Saunders, C. (2019). How communication between cancer patients and their specialists affect the quality and cost of cancer care. Supportive care in cancer, 27, 4575-4585.

[20] Suganthi, S. T., Ayoobkhan, M. U., Kumar, K., Bacanin, N., K, V., Štěpán, H., & Pavel, T. (2022). Deep learning model for deep fake face recognition and detection. PeerJ Computer Science, 8, e881.

[21] Taiwo Olaleye, O. A. (2021). Evaluation of image filtering parameters for plant biometrics improvement using machine learning. Soft Computing and its Engineering Applications: Second International Conference, icSoftComp 2020 (pp. 1–12). Anand: Springer Singapore.

[22] Thom, B., & Benedict, C. (2019). The impact of financial toxicity on psychological well-being, coping self-efficacy, and cost-coping behaviors in young adults with cancer. Journal of adolescent and young adult oncology, 8(3), 236-242.

[23] Wang, M., Li, X., Chen, L., & Chen, H. (2023). Medical machine learning based on multiobjective evolutionary algorithm using learning decomposition. Expert Systems with Applications, 216. doi:https://doi.org/10.1016/j.eswa.2022.119450

[24] Weinberg, R. A., & Weinberg, R. A. (2006). The biology of cancer. WW Norton & Company.

[25] Woolf, S. H. (2008). The meaning of translational research and why it matters. Jama, 299(2), 211-213.