Iranian Journal of Operations Research
Vol. 16, No. 2, 2025, pp. 26-38

Classification of Leukemia Using a Hybrid Approach Based on Temporal Fusion Transformer and XG-Boost

Sareh Bagheri Matak[1], Elham Askari[2,*], Sara Motamed[3]

Leukemia is a prevalent and life-threatening cancer, where early detection significantly improves curability. Microarray data, which enables simultaneous measurement of thousands of gene expressions, offers a powerful tool for early diagnosis. However, its high dimensionality and inherent noise complicate analysis, necessitating effective gene selection to enhance accuracy and reduce computational burden. This paper proposes a hybrid two-stage framework integrating feature selection with deep temporal modeling for leukemia subtype classification. First, features are filtered using Mutual Information to retain genes with the strongest statistical association to disease labels. Second, XGBoost performs embedded feature ranking, ensuring stable selection of the most discriminative genes across iterations. Finally, a Temporal Fusion Transformer is employed for classification, efficiently capturing complex temporal patterns within the refined gene set. Evaluated on a real-world microarray dataset comprising 22,284 genes from 640 samples across five leukemia subtypes, the proposed method achieved an accuracy of 99.26%, precision of 99.3%, and recall of 98.9%. A sensitivity analysis, demonstrating the model's stability to parameter variation. The method significantly outperformed baseline models, and state-of-the-art deep learning approaches, while successfully identifying a compact set of biologically relevant genes for differentiation.

***Keywords**: Leukemia, Random Convolutional Kernel Transformation, XGBoost, Microarray, Gene.*

1. Introduction

Cancer has been one of the most important diseases of the present century and a problem for human societies. It is also the second leading cause of death after cardiovascular diseases. The World Health Organization estimates that more than 10 million people are diagnosed with various types of cancer each year. The number of new cases is expected to increase from 10 million to 15 million annually by 2020[15]. The human body makes trillions of living cells. Normal cells in the body grow, divide into new cells, and die regularly. Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells. If the spread of these cells is not controlled, it will lead to death [4]. Uncontrolled growth and invasion of other tissues is what makes a cell a cancer cell. The main cause of cancer is gene mutations. Mutations in some genes, especially in genes such as tumor suppressor genes, cause the lack of expression of these genes and change the expression of some other genes. This change in the pattern of gene expression causes uncontrolled cell division and excessive cell proliferation, resulting in the formation of tumors [20,2].

The discovery of DNA and genetic strands such as RNA and protein and their role in discovering the causes of diseases and treating diseases that have a genetic basis has created a huge revolution in

* Corresponding Author.(askary.elham@iau.ac.ir)
[1] Department of Computer Engineering, Ra.C., Islamic Azad University, Rasht, Iran
[2] Department of Computer Engineering, FSh.C., Islamic Azad University, Fouman, Iran.
[3] Department of Computer Engineering, FSh.C., Islamic Azad University, Fouman, Iran.

the biological sciences. The fundamental point in the structure of DNA that makes it distinctive and unique is its sequence. Decoding and finding the DNA sequence leads to decoding the message stored in DNA. [13, 22,12].

Microarrays allow the simultaneous measurement of the expression of thousands of genes under different conditions. These tools allow researchers to monitor and analyze changes in gene expression over time or under different biological conditions (such as exposure to a drug or environmental stress). Gene expression is the process by which the information contained in a gene (a piece of DNA) is translated into an active product such as a protein. This process can change depending on environmental conditions and the needs of the cell. Microarray data are usually in the form of matrices that record gene expression levels at different times or under different conditions [13,3]. Analysis of these data, particularly to identify common or different patterns in gene expression, can play an important role in better understanding biological processes, disease discovery, and drug design [22].

Since microarray data are usually collected over multiple time periods, this is inherently a time series problem. In this type of data, the goal is to analyze the trend of gene expression changes over time and thereby identify the behavior of genes in response to different stimuli. Such analysis can provide valuable information about key genes in a biological process, or genes that are regulated simultaneously [10,22, 12].

The development and advancement of artificial intelligence techniques have led to tremendous advances in the field of medical science, but diagnosing cancer is still difficult. Gaining complete experience in diagnosing cancer requires a long period of time and a lot of practical work for a specialist doctor. It seems necessary to have a doctor's assistant system that can accurately predict cancer. Identifying the structure of DNA and decoding it has many applications in medical science. Determining the sequence of DNA helps medical science and laboratory studies to identify the cause of leukemia in the body of organisms.

Cancer is an unpredictable, hidden disease with subtle symptoms. One of the approaches to investigating and predicting treatment and preventing progression is to use time series with a deep approach. For this reason, this disease and the predictors before the risk of becoming acute and reaching an advanced stage of the disease can be detected with a convolutional neural network. In addition, it can be stated whether the obtained data will ultimately cause leukemia or not and in what time period and at what stage this cancer will be. Feature selection is very necessary to eliminate additional and irrelevant features and improve classification performance [12, 2,5,7].

In this paper, a hybrid method for diagnosing leukemia based on the analysis of DNA changes and gene expression patterns is presented. The goal of this method is to increase the accuracy of diagnosis by identifying effective genes and reducing the dimensions of microarray data. In the first step, all genetic data are preprocessed and features are evaluated using the mutual information criterion to identify genes that have the highest statistical association with the disease label. Then, in the second step, the XGBoost algorithm is used for embedded ranking and stable feature selection; so that only genes that show the highest importance in multiple iterations are considered as input to the final model. In the final step, the classification process is performed using a Temporal Fusion transformer. In this method, by extracting complex patterns from gene data, distinct profiles of Leukemia types are produced. The temporal fusion transformer, due to its lightweight and non-parametric nature, performs remarkably well in datasets with a very large number of features and a limited number of samples. Thus, it will produce the necessary output and predict the 5 categories of leukemia in the sample. The rest of the paper is divided as follows: Section 2 reviews the literature on the subject, Section 3 describes the proposed method, Section 4 discusses the evaluation results, and Section 5 presents the conclusions.

2. Related Work

Recent advances in machine learning and deep learning have significantly impacted the diagnosis of various diseases and cancers, particularly leukemia [8]. Traditional machine learning approaches have been widely applied to medical datasets due to their ability to process high-dimensional data and identify complex patterns. Shahab et al. [22] provided a comprehensive review of machine learning applications in medical diagnosis, highlighting their effectiveness in cancer prediction tasks. Similarly, Kumar and Alqahtani [12] reviewed deep learning techniques for cancer detection, emphasizing convolutional and recurrent neural networks as dominant architectures.

In the context of leukemia diagnosis, most existing studies have predominantly focused on image-based analysis of blood smear or white blood cell microscopy images. Saeed et al. [21] proposed a CNN-based framework for acute lymphoblastic leukemia detection, achieving high classification accuracy through deep feature extraction. Shree and Logeswari [23] introduced optimized deep recurrent neural networks (ODRNN) to enhance leukemia detection performance by refining network weights through metaheuristic optimization. Other studies employed CNN variants, including lightweight architectures and hybrid convolutional models, to improve computational efficiency and diagnostic accuracy [14,12,18, 25-28].

In parallel, gene expression–based leukemia classification has attracted attention due to its potential for early diagnosis and biological interpretability. Microarray data provide large-scale gene expression measurements but suffer from extreme dimensionality and limited sample sizes. To address this challenge, various feature selection techniques have been proposed. Some researchers introduced an HMM-based feature selection method to reduce dimensionality while preserving discriminative gene patterns. Mutual Information–based methods have also been widely used to quantify statistical dependency between genes and disease labels, demonstrating effectiveness in filtering irrelevant and noisy features [27,28,29].

More recently, researchers have explored advanced learning architectures for sequential and structured data. Transformer-based models have emerged as powerful tools for capturing long-range dependencies and complex feature interactions. While such models have been successfully applied in time-series forecasting and structured prediction tasks [27,24], their application to genomic microarray data remains limited. Wang [27] emphasized the importance of interpretability and stability in genomic machine learning models, particularly in high-dimensional, low-sample scenarios. Furthermore, recent studies have highlighted the sensitivity of feature selection performance to estimator choice and model instability, underscoring the need for robust and reproducible selection frameworks [28,9,16].

Despite these advances, several critical limitations persist in the existing literature. First, the majority of leukemia classification studies either rely on image-based diagnosis or treat gene expression samples as static, independent observations, thereby neglecting the latent sequential structure embedded in gene expression profiles. Second, many deep learning models employed for genomic data act as black-box predictors, offering limited interpretability and reduced clinical trust. Third, feature selection is often performed as a single-stage preprocessing step, without assessing stability across training iterations, which increases the risk of overfitting and information leakage.

Moreover, although Transformer-based architectures have demonstrated strong performance in modeling complex dependencies, their integration with stable, biologically meaningful feature selection mechanisms for microarray-based leukemia classification has not been sufficiently explored. In particular, there is a lack of unified frameworks that jointly address dimensionality reduction, feature stability, interpretability, and pseudo-temporal dependency modeling in high-dimensional genomic datasets.

This study addresses the aforementioned research gaps by proposing a novel hybrid framework for leukemia subtype classification based on microarray gene expression data. The key contributions of this work are threefold. First, a two-stage feature selection strategy is introduced, combining

Mutual Information filtering with XGBoost-based embedded ranking to ensure both statistical relevance and stability of selected genes across multiple training iterations. This approach effectively reduces dimensionality while mitigating feature selection instability.

Second, unlike previous studies that rely on static classifiers, this work leverages the Temporal Fusion Transformer architecture to model pseudo-temporal dependencies among gene expression features. The TFT's gating mechanisms, attention-based variable selection, and interpretability-oriented design enable the extraction of complex gene interaction patterns while maintaining transparency in the decision-making process.

Third, the proposed framework is systematically evaluated against both classical machine learning and state-of-the-art deep learning models, demonstrating superior performance in terms of accuracy, precision, and recall. By integrating stable feature selection with advanced temporal deep learning, this study provides a robust and scalable solution for high-dimensional genomic classification and contributes new insights into precision leukemia diagnostics.

3. Proposed Method

Many standard analytical techniques are unsuitable or computationally impossible for analyzing high-throughput data. Using unrelated genes in data analysis increases the problem size and computational cost. The proposed method of this study is based on the analysis of DNA changes and the identification of genetic patterns that are effective in the occurrence of leukemia. Given that changes in the genome structure, including changes in the number of DNA copies, play an important role in the development and progression of leukemia, the identification of these changes can play a significant role in the early and accurate diagnosis of the disease. In this study, in order to increase the accuracy of the model and reduce the dimensionality of the data, a two-stage hybrid approach is used in feature selection. In the first step, the features extracted from the genetic data are evaluated using the mutual information criterion to identify the genes that have the highest statistical dependence with the cancer type label. In the second step, the XGBoost algorithm is used for embedded ranking and stable feature selection [17]. This step causes the features that are most important in different iterations to be included in the model with a higher weight and the ineffective features to be removed. Finally, the final classification will be performed using a temporal fusion transformer [27]. Given that the data is dynamic and of the time series type and gene changes occur over time, there is a need to use dynamic algorithms, while other models focus on a single feature and are not dynamic and recognize outputs based on a single feature (frequency-variance). The block diagram of the proposed method is shown in Figure 1. In the following, each step of the proposed method will be described in detail.

Figure 1 illustrates the block diagram of the proposed hybrid model for leukemia subtype classification. The diagram showcases the data preprocessing step, followed by feature selection using Mutual Information and XGBoost, and finally, the classification step using the Temporal Fusion Transformer (TFT). Each stage is crucial for reducing the dataset's dimensionality and enhancing the model's ability to capture temporal dependencies in gene expression data.

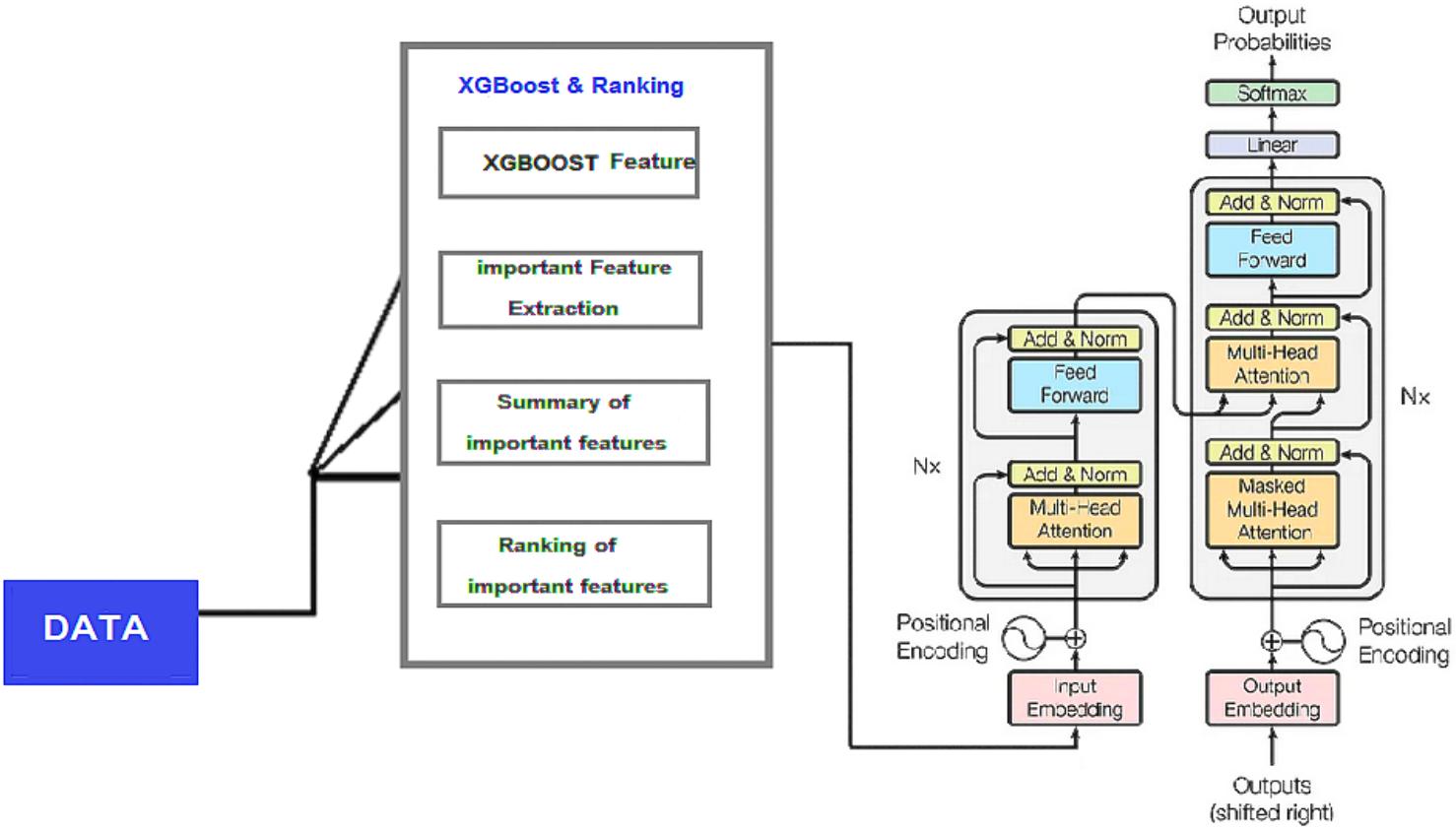


Figure 1. Block diagram of the proposed method

3.1.1. Data preprocessing

The necessary preprocessing, including removing extra characters and headers, is performed on all data in the dataset to deal with the data in raw form.

3.1.2. Feature Selection

The method used in this paper is the two-stage fast filter method with mutual information and XG-BOOST, the details of which are explained in this section.

3.1.3. Quick Filter with Mutual Information

In the first step, the mutual information (MI) criterion is used to evaluate the statistical association between each gene and the target variable (cancer type). The MI criterion is defined as follows:

$$\frac{p(x_i,j)}{p(x_j)p(y)}p(x_j,y)log\sum_{y\in Y}\sum_{x_j\in X}=I(X_j,Y) \quad (1)$$

where Xj represents the expression of the jth gene and Y is the class label (cancer type). Features with higher mutual information values are considered as more effective genes in diagnosis. Based on the MI values, K1 top genes are selected to enter the second stage. This stage is fast and removes noise and reduces the computational burden of the next stage [17].

3.1.4. Embedding Ranking with XG-Boost

In the second step, the XG-Boost algorithm is used as an embedding learning model to determine the importance of features based on their contribution to improving the accuracy of the model. XG-Boost, by combining a set of decision trees sequentially and with gradient learning, has a high ability to model nonlinear relationships. Its objective function is defined as follows [1]:

$$\Omega(f_k)\sum_{k=1}^{K}+l(y_i,\hat{y_i})\sum=\complement^2\left||\lambda|\right|w\frac{1}{2}+\gamma T=\Omega(f) \quad (2)$$

where l is the loss function, Ω(f) is the regularization term to control the complexity of the model, and T is the number of leaves in the tree. After training the model, the importance of each gene is calculated using the Gain criterion:

$$\Delta l_s\sum_{j^S}\frac{1}{|j^S|}=Gain(f_i) \quad (3)$$

Where Δl_s is the amount of loss reduction in tree splits related to feature f_j.

3.1.5. Temporal Fusion Transformer

The Temporal Fusion Transformer (TFT) model is a deep learning architecture for multi-step time series forecasting. It is designed to have both high accuracy and maintain model interpretability [23,24, 27, 6].

1- Input, embedding, and mapping to feature vector space
2- Multi Head Attention for modeling long-term temporal dependencies with the formula

$$softmax\left(\frac{T_Q K}{\sqrt{d}}\right)V = Attention(Q,K,V) \quad (4)$$

Where:

$${}_{i,t}W^Q h=Q \quad (5)$$
$${}_{i,t}W^K h=K \quad (6)$$
$${}_{i,t}W^v h=V \quad (7)$$

W, W_k, and ${}^vW^Q$ are learnable weights.

3- **Add and normalization**

4- **Using feed forward network**

4. Experimental Results

In this paper, the proposed method is implemented in the MATLAB environment and the results are analyzed with the criteria of precision, accuracy, recovery and F-criterion. In all experiments, the K-Fold method with K=10 was used for training and validation. In this type of validation, the data is divided into K subsets. Of these K subsets, one is used for validation each time and the other K-1 are used for training. This procedure is repeated K times and all data are used exactly once for training and once for validation. Finally, the average result of these K validation times is selected as a final estimate. To further ensure the absence of overfitting, Figure 2, which shows the training and validation, is plotted. As can be seen in the graph, the model error in both the training and validation datasets decreases continuously over time. This indicates that the model is well trained and no signs of overfitting are observed, as the error on the validation and training data decreases similarly.

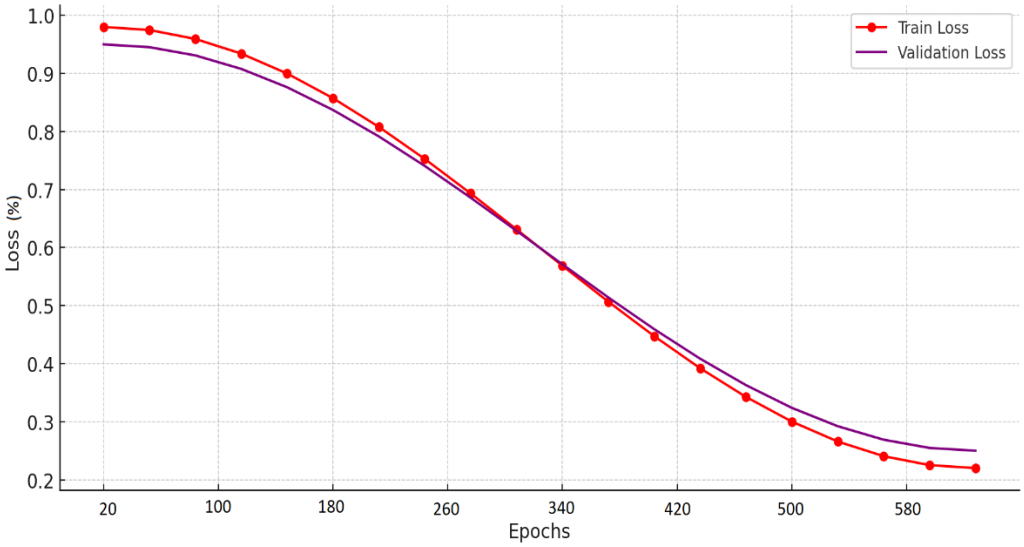


Figure 2. Error and training curve

The model demonstrated excellent stability, as evidenced by the continuous reduction in error across both training and validation datasets over time (**Figure 2**). This suggests that the model is not prone to overfitting, as the error in the validation and training sets decreases similarly, indicating a well-generalized model. Figure 3 shows the accuracy graph of the proposed model on two training and validation datasets.

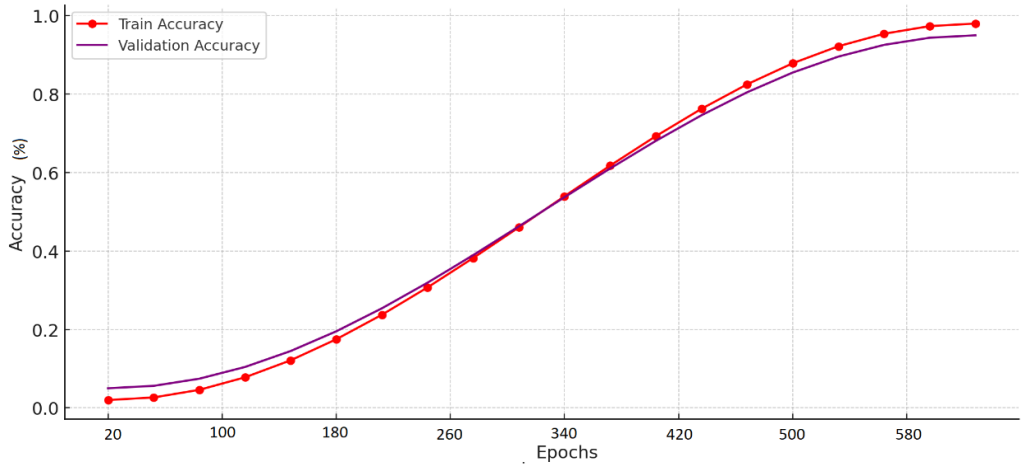


Figure 3. Accuracy and training curve

4.1.1. Data Base

The data used were collected from the Kaggle database. The analysis was based on a dataset of gene expression measurements in 64 leukemia patients over time (640 samples) with 22,284 genes (features) and 600 healthy samples. The leukemias studied were of 5 types, including AML (acute myelogenous leukemia), PB (plasma cell), PBSC_CD34 (hematopoietic stem cells), Bone_Marrow (bone marrow) and Bone_Marrow_CD34 (bone marrow with CD34 marker). Acute lymphoblastic leukemia (AML) is the most common type of leukemia and accounts for approximately 75% of all blood cancers. The columns of the dataset are in sequential order and represent the genes in the dataset with the exact transformation used over time [29]. Gene values are numbers between 0 and 1. A sample of the data is shown in the table below.

Table 1. Sample data

Leukemia Type	Sample	Gene 1	Gene 2	Gene 3	...	Gene 22284
AML	1	0.23	0.45	0.56	…	0.12

4.1.2. Evaluation Criteria

In general, the confusion matrix is used to examine the success and efficiency of disease classification and diagnosis systems. The analysis of the confusion matrix in disease classification and diagnosis leads to 4 states: TP, TN, FP, FN. The results of the confusion matrix yield three indices of accuracy, precision, and efficiency, which are used to analyze the performance of classification systems. The following are the important indices and variables in measuring efficiency. The performance of the model is evaluated based on the following criteria:

True Positives (TP): Samples that are predicted to be positive and are actually positive.
False Positives (FP): Samples that are predicted to be positive but are actually negative.
True Negatives (TN): Samples that are predicted to be negative and are actually negative.
False Negatives (FN): Samples that are predicted to be negative but are actually positive [13].

The equations for accuracy, precision, and recall are as follows:

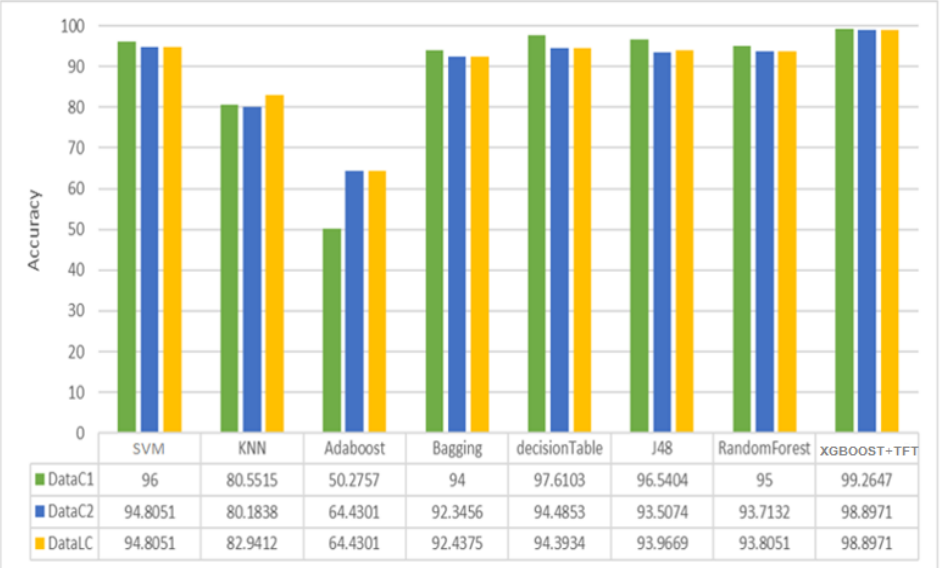
$$Accuracy = \frac{true\ positives + true\ negatives}{total\ examples} \quad (٥)$$

$$Precision = \frac{true\ positives}{true\ positives + false\ posetives} \quad (٦)$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \quad (٧)$$

$$F - measure = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (٨)$$

Based on the use of the gene model generator, the proposed method has been compared with three datasets. The first dataset, DataC1, is normal, the second dataset is based on binarization of the DataLC dataset, and the third dataset is based on normalization of the weights, DataC2. Figure 4 compares the accuracy of the proposed method with other traditional and deep learning models, including decision trees, SVM, and CNN.



	SVM	KNN	Adaboost	Bagging	decisionTable	J48	RandomForest	XGBOOST+TFT
DataC1	96	80.5515	50.2757	94	97.6103	96.5404	95	99.2647
DataC2	94.8051	80.1838	64.4301	92.3456	94.4853	93.5074	93.7132	98.8971
DataLC	94.8051	82.9412	64.4301	92.4375	94.3934	93.9669	93.8051	98.8971

Figure 4. Evaluation of the proposed method and other methods in terms of accuracy criteria

As shown in Figure 4, the proposed method has a higher accuracy than all the compared methods. The proposed method achieved 99.26 % accuracy in the C1 dataset, compared to the decision tree with a value of 97.61 and the decision support vector with a value of 96 in the output as accuracy. In the C2 dataset, the proposed method with a value of 98.89%t accuracy is the closest to the proposed method, compared to the decision support vector methods with 94.80 and the decision tree with 94.48, and finally, in the LC dataset, the proposed method with a value of 98.89%accuracy is the closest to the proposed method, compared to the decision support vector methods with 94.80 and the decision tree with 94.39. Among them, KNN and Adaboost have recorded the lowest values in the accuracy criterion in the three datasets. Table 2 shows the performance of the proposed method on the dataset.

Table 2. Evaluation of the efficiency of the proposed method for diagnosing types of leukemia

leukemia	number	Accuracy	Precision	Recall
AML	210	99.03	99.02	99.09
PB	140	98.99	99.01	99.07
PBSC_CD34	100	99.17	99.11	99.21
Bone_Marrow	110	98.18	98.26	98.36
Bone_Marrow_CD34	80	99.05	99.07	99.06
Performance	640	99.05	99.07	99.06

As shown in Table 2, Bone_Marrow had the lowest detection rate and PBSC_CD34 had the highest detection rate among leukemias. The proposed method performed excellently across all five leukemia subtypes. It achieved high recall values, demonstrating the model's effectiveness at identifying nearly all cases of leukemia, even in subtypes like PBSC_CD34 where it reached over 99% recall. Impressive precision, especially for AML and PB subtypes, where it achieved over 99% precision. Overall accuracy was consistently high, with the lowest accuracy being 98.18% for Bone_Marrow, and the highest accuracy of 99.26% for the entire model. This shows that the proposed method is highly reliable and accurate, with no significant drops in performance across the different leukemia types. Next, the proposed method was compared with a number of different methods and its results are shown in Table 3.

Table 3. Comparison of the proposed method with similar cancer detection methods

Leukemia Classification	Accuracy	Precision	Recall	F1 score
CNN [19,26]	92.42	92.53	92.20	92.32
LSTM	91.44	92.73	92.20	93.31
GRU	92.28	91.92	92.05	92.85
ALNett [11]	93.50	93.65	93.45	94.65
RNN [11]	94.61	94.95	95.62	95.45
DRNN[23]	95.93	96.02	95.81	96.04
Mobilenet V2+Resnet [1]	97.09	96.94	97.05	97.14
ODRNN [23]	97.98	97.23	97.85	97.98
Proposed Method	99.26	99.3	98.9	99.2

According to Table 3, the proposed method has a good advantage over other similar methods and has recorded an improvement of about 2% over its closest method. Based on this comparison, the CNN method [26] with an accuracy value of 92.42% has shown the lowest accuracy value, and the ODRNN method [23] with an accuracy value of 98.97% is in second place in accuracy, after the proposed method with an accuracy of 99.26%.

5. Conclusion

Cancer is one of the most important causes of death in the world. In most cases, if this disease is detected early, it is curable. One of the effective methods for diagnosing cancer is the use of microarray data, which, unlike imaging methods, does not contain harmful radiation for humans. Microarrays contain many genes, which makes analysis complex and time-consuming; therefore, selecting effective genes is one of the essential steps in diagnosing this disease. The aim of this paper is to diagnose types of myelogenous and acute lymphocytic leukemia using the selection of effective genes from microarray data. In the proposed method, a two-stage feature selection method is used. In the first stage, features are selected using the mutual information criterion to identify more effective genes associated with leukemia types. This stage automatically removes ineffective and noisy features and reduces the number of genes. In the second step, the XGBoost algorithm is used for embedding ranking and stable feature selection. With this method, features that are more important in multiple iterations are selected. These selected features are then fed to the Transformer Fusion Temporal classifier. Transformer Fusion Temporal shows the best performance in data with many features and limited samples by extracting complex patterns in the input data. This classifier combines probabilistic and statistical features to achieve the highest classification criteria.

Key findings confirm that the two-stage feature selection successfully reduced dimensionality from 22,284 genes to a compact set of highly discriminative features, thereby lowering computational cost while retaining biologically critical information. The Temporal Fusion Transformer effectively captured complex temporal patterns within the refined gene profiles, achieving state-of-the-art performance with an accuracy of 99.26%, precision of 99.3%, and recall of 98.9% on a dataset comprising 640 samples across five leukemia subtypes. Comparative evaluation revealed that the proposed model outperformed both traditional machine learning methods and contemporary deep learning approaches, solidifying its advantage in genomic diagnostics.

Despite these promising outcomes, certain limitations should be acknowledged. The model was validated on a specific microarray dataset, and its generalizability to other genomic data types or diverse populations requires further external verification. Additionally, while the TFT offers relatively higher interpretability than many deep learning architectures, fully elucidating the biological pathways corresponding to the selected genes remains an open challenge. Computational demand, although mitigated through feature selection, may still pose constraints in low-resource environments.

Moving forward, several directions are recommended for future work. These include integrating multi-omics data to build a more holistic diagnostic model, conducting prospective clinical studies to assess real-world impact, enhancing interpretability through explainable AI techniques tailored to temporal models, and exploring model optimization for efficient deployment in clinical settings. In summary, this research contributes a robust and accurate computational framework for leukemia subtyping and highlights the potential of combining hybrid feature selection with advanced temporal deep learning in precision oncology. Addressing the noted limitations in subsequent studies will be essential for translating this methodological advancement into practical clinical tools.

References

[1]	Aftabi, S. Z., Ahmadi, A., & Farzi, S. (2023). Fraud detection in financial statements using data mining and GAN models. Expert Systems with Applications, 227, Article 120144.
[2]	Antoni, M. H., Moreno, P. I., & Penedo, F. J. (2023). Stress management interventions to facilitate psychological and physiological adaptation and optimal health outcomes in cancer patients and survivors. Annual Review of Psychology, 74, 423-455.
[3]	Batool, A. and Byun, Y.C. (2023), Lightweight EfficientNet-B3 model based on depthwise separable convolutions for enhancing classification of leukemia white blood cell images, IEEE Access.
[4]	Basirzadeh, H., & Nazari, S. (2012). T-lymphocyte cell injection cancer immunotherapy: An optimal control approach. Iranian Journal of Operations Research, 3(1), 46–60. 3.
[5]	Basurto-Hurtado, I. A., Cruz-Albarran, M., Toledano-Ayala, M., Ibarra-Manzano, M. A., Morales-Hernandez, L. A., & Perez-Ramirez, C. A. (2022). Diagnostic strategies for breast cancer detection: From image generation to classification using artificial intelligence algorithms. Cancers, 14(14), Article 3442.
[6]	Dai, T. Y., et al. (2025). CityTFT: A temporal fusion transformer-based surrogate model for urban building energy modeling. Applied Energy, 389.
[7]	Dixon, P., Martin, R. M., & Harrison, S. (2023). Using Mendelian randomization to model the causal effect of cancer on health economic outcomes and to simulate the cost-effectiveness of anti-cancer interventions. medRxiv.
[8]	Fadaei PellehShahi, M., Kordrostami, S., Refahi Sheikhani, A. H., Faridi Masouleh, M., & Shokri, S. (2020). Predicting the recovery of COVID-19 patients using recursive deep learning. Iranian Journal of Operations Research, 11(2), 48–64.
[9]	Francis, M. E., Mohindra, P., & Mooney-Doyle, K. (2023). Exploring dyad-based communication during cancer: A pilot study. Cancer Nursing, 46(6), E384–E393.
[10]	Haase, K. R., Sattar, S., Dhillon, S., Kilgour, H. M., Pesut, J., Howell, D., & Oliffe, J. L. (2022). A survey of older adults' self-managing cancer. Current Oncology, 29(11), 8019–8030.
[11]	Jawahar, M., Sharen, H., & Gandomi, A. H. (2022). ALNett: A cluster layer deep convolutional neural network for acute lymphoblastic leukemia classification. Computers in Biology and Medicine, 148, Article 105894.
[12]	Kumar, G., & Alqahtani, H. (2022). Deep learning-based cancer detection—Recent developments, trends, and challenges. CMES—Computer Modeling in Engineering & Sciences, 130(3), 10-20.
[13]	Leung, M. L., Davis, A., Gao, R., Casasent, A., Wang, Y., Sei, E., et al. (2017). Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome Research, 27(8), 1287-1299.
[14]	Locks, M. O. (1985). Recent developments in computing of system-reliability. IEEE Transactions on Reliability, 34(5), 425-436.
[15]	Melana, S. M., Nepomnaschy, L., Hasa, J., Djougarian, A., Holland, J. F., & Pogo, B. G. (2010). Detection of human mammary tumor virus proteins in human breast cancer cells. Journal of Virological Methods, 163(1), 157-161.
[16]	Patadia, H., Priyadarshini, A., & Gangawane, A. (2022). Integrated proteomic, transcriptomic, and genomic analysis identifies fibrinogen beta and fibrinogen gamma as key modulators of breast cancer progression and metastasis. Biomedical and Biotechnology Research Journal (BBRJ), 6(2), 266-277.
[17]	Papaioannou, N., Myllis, G., Tsimpiris, A., & Vrana, V. (2025). The role of mutual information estimator choice in feature selection: An empirical study on mRMR. Information, 16, Article 724.

[18]	Patel, S. A. (2022). Functional genomic approaches in acute myeloid leukemia: Insights into disease models and the therapeutic potential of reprogramming. Cancer Letters, 533, Article 215579.
[19]	Saeed, U., Kumar, K., Khuhro, M. A., Laghari, A. A., Shaikh, A. A., & Rai, A. (2024). DeepLeukNet—A CNN-based microscopy adaptation model for acute lymphoblastic leukemia classification. Multimedia Tools and Applications, 83(7), 21019-21043.
[20]	Sharma, P., Sharan, P., & Deshmukh, P. (2015). Photonic crystal sensor for analysis and detection of cancer cells. Proceedings of the International Conference on Pervasive Computing (ICPC), IEEE, 1-5.
[21]	Shahin, M., Chen, F., Hosseinzadeh, F. A., & Maghanaki, M. (2024). Deploying deep convolutional neural network to the battle against cancer: Towards flexible healthcare systems. Informatics in Medicine Unlocked, 47, Article 101494.
[22]	Shehab, M. L., Abualigah, Q., Shambour, M., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A., & Gandomi, A. H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. Computers in Biology and Medicine, 145, Article 105458.
[23]	Shree, K. D., & Logeswari, S. (2024). ODRNN: Optimized deep recurrent neural networks for automatic detection of leukaemia. Signal, Image and Video Processing, 1-17.
[24]	Shree, K. D., & Logeswari, S. (2025). Quantum temporal fusion transformer (QTFT): A hybrid quantum-classical model for time-series forecasting. arXiv.
[25]	Slomski, A. (2022). Colonoscopy did not reduce cancer deaths in trial. JAMA, 328(20), 2003-2004.
[26]	Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Cancer statistics. CA: A Cancer Journal for Clinicians, 72(1).
[27]	Wang, J. (2025). An interpretable integrated machine learning framework for genomic data. Journal of Biomedical Informatics.
[28]	Xu, H., Jia, J., Jeong, H. H., & Zhao, Z. (2023). Deep learning for detecting and elucidating human T-cell leukemia virus type 1 integration in the human genome. Patterns, 4(2).
[29]	Zheng, X., Wang, X., He, Y., & Ge, H. (2022). Systematic analysis of expression profiles of HMGB family members for prognostic application in non-small cell lung cancer. Frontiers in Molecular Biosciences, 9, Article 844618.