Karadeniz Fen Bilimleri Dergisi, 15(3), 1046-1069, 2025. DOI: 10.31466/kfbd.1597865

KFBD

Karadeniz Fen Bilimleri Dergisi

The Black Sea Journal of Sciences

ISSN (Online): 2564-7377 https://dergipark.org.tr/tr/pub/kfbd

Araştırma Makalesi / Research Article

Discovery of Marker Genes in Adult T Cell Leukemia (ATL) Pathogenesis with Machine Learning Models and Performance Comparison

Sabire KILICARSLAN[1], Sait Can YUCEBAS[2*]

Abstract

Hematologic cancers are often diagnosed after symptoms become apparent, which can make it difficult to control the disease and implement effective treatment strategies. Studying gene expression profiles is vital for early diagnosis and the development of treatment strategies for hematologic cancers such as T-cell leukemia. The motivation of this study is to reveal the molecular mechanisms in the pathogenesis of this disease by comparing the whole gene expression profile in Adult T-cell Leukemia (ATL) cells and CD4+T cells of healthy individuals. For this aim, several machine learning algorithms, Naive Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, C4.5, Logistic Regression, Linear Discriminant Analysis and Artificial Neural Network algorithms were used. Their performance was compared on the GSE33615 dataset by using 5-fold cross validation with stratified sampling. Among these, Artificial Neural Network stood out with an AUC of 0.98 and an F1 score of 0.93. It was followed by SVM with an AUC of 0.97 and 0.957 F1 score. In addition to performance comparison, information gain ratio, SHAPLEY metric and correlation values were calculated for the detection of genes causing ATL. Among the models, the three with the highest performance (ANN, SVM, RF) were selected, and the top ten most significant genes were identified for each. Considering the intersection of these gene sets, ZSCAN18, PLK3, and NELL2 were found to be associated with the related disease. These genes may contribute to Adult T-cell Leukemia pathogenesis through their roles in cell cycle regulation, transcriptional control, and oncogenic signaling. Further investigation is needed to clarify their precise molecular mechanisms in the related disease.

Keywords: Adult T-cell Leukemia (ATL), Microarray study, Machine learning, Variable importance.

Erişkin T Hücreli Lösemi (ATL) Patogenezindeki Marker Genlerin Makine Öğrenmesi Modelleri ile Keşfi ve Performans Karşılaştırması

Öz

Hematolojik kanserler genellikle semptomlar belirginleştikten sonra teşhis edilir ve bu durum hastalığın kontrol altına alınmasını ve etkili tedavi stratejilerinin uygulanmasını zorlaştırabilir. Özellikle T hücreli lösemi gibi hematolojik kanserlerde, gen ekspresyon profillerinin incelenmesi, erken tanı ve tedavi stratejilerinin geliştirilmesinde hayati öneme sahiptir. Bu çalışma, Yetişkin T hücreli Lösemi (ATL) hücrelerinde ve sağlıklı bireylerin CD4+T hücrelerindeki tüm gen ekspresyon profilini karşılaştırarak, bu hastalığın patogenezindeki moleküler mekanizmaları farklı makine öğrenme yöntemleri ile ortaya çıkarma motivasyonu ile gerçekleştirilmiştir. Naive Bayes, K-En Yakın Komşu, Destek Vektör Makinesi, Rassal Orman, C4.5, Lojistik Regresyon, Doğrusal Diskriminant Analizi ve Yapay Sinir Ağları algoritmalarının karar performansları, GSE33615 veri seti üzerinde tabakalı örnekleme ile 5 katlı çapraz doğrulama yöntemi kullanılarak karşılaştırılmıştır. Bunlar arasında Yapay Sinir Ağı 0,98 AUC ve 0,93 F1 skoru ile öne çıkmıştır. Onu, 0.97 AUC ve 0.957 F1 skoru ile SVM takip etmiştir. Performans karşılaştırmasına ek olarak, ATL'ye neden olan genlerin tespiti için bilgi kazanç oranı, SHAPLEY metriği ve korelasyon değerleri hesaplanmıştır. Her model için en yüksek öneme sahip ilk on gen belirlenmiştir. Modeller tarafından önerilen genlerin kesişim kümesi dikkate alındığında, ZSCAN18, PLK3 ve NELL2 genlerinin ilgili hastalık için ilişkili olduğu bulunmuştur. Bu genler, hücre döngüsü düzenlenmesi, transkripsiyonel kontrol ve onkojenik sinyal iletimi üzerindeki rollerine bağlı olarak Erişkin T-hücreli Lösemi patogenezine katkıda bulunabilir. Bu genlerin moleküler rollerinin daha iyi anlaşılabilmesi için ileri araştırmalara ihtiyaç duyulmaktadır.

Anahtar Kelimeler: Yetişkin T-hücreli Lösemi (ATL), Mikroarray çalışması, Makine öğrenmesi, Değişken önemi.

[1]Çanakkale Onsekiz Mart University, Department of Medical System Biology, Faculty of Medicine, Çanakkale, Türkiye, sabire.kilicarslan@gmail.com
[2]Çanakkale Onsekiz Mart University, Department of Computer Engineering, Faculty of Engineering, Çanakkale, Türkiye, can@comu.edu.tr

[*]Sorumlu Yazar/Corresponding Author

Geliş/Received: 07.12.2024 **Kabul/Accepted:** 09.09.2025 **Yayın/Published:** 15.09.2025

Kılıcarslan, S., Yucebas, S.C. (2025). Discovery of Marker Genes in Adult T Cell Leukemia (ATL) Pathogenesis with Machine Learning Models and Performance Comparison. *Karadeniz Fen Bilimleri Dergisi*, 15(3), 1046-1069.

1. Introduction

Adult T-cell leukemia/lymphoma (ATL) is a highly aggressive disease caused by human T-cell leukemia virus type I (HTLV-1) with an extremely bad prognosis (Ishitsuka and Tamura, 2014; Uchiyama et al., 1977). The median overall survival of the aggressive subtypes, including the acute and lymphoma types (about 60% of cases), is only sketchy to ten months (Katsuya et al., 2015). Even those initially diagnosed in indolent forms, such as smoldering and chronic subtypes, usually progress to aggressive disease within a year (Takasaki et al., 2010). HTLV-1 infection is estimated to affect 5 to 20 million people worldwide (Gessain and Cassar, 2012), with higher prevalence in regions such as southwestern Japan, the Caribbean Basin and central Africa (Ishitsuka and Tamura, 2014). While HTLV-1 infection usually leads to a lifelong carrier state, less than 5% of infected individuals die from HTLV-1-associated leukemia. ATL leukemogenesis involves the accumulation of multiple genetic abnormalities in HTLV-1-infected cells, a complex process.

The diagnosis of ATL is usually made by detection of HTLV-1 antibodies in the light of clinical signs such as lymph node enlargement, skin lesions, hypercalcemia and subsequent confirmation of HTLV-1 proviral DNA by PCR (Cook et al., 2021). Early diagnosis of ATL is critical, as the aggressive progression of the disease often results in mortality.

While recent studies have introduced various ML models for the diagnosis of ATL, these efforts often focus solely on predictive performance using a single method, such as deep learning (Kılıçarslan and Pacal, 2023; Xu et al., 2023), support vector machines (Chong et al., 2020), random forests (Faiz et al., 2024), or decision trees (Eckardt et al., 2020). However, very few studies emphasize the identification of the underlying genome profiles (Abass and Adeshina, 2021; Stricker et al., 2017), which are essential for understanding disease mechanisms and developing effective treatment strategies. Furthermore, model interpretability and biological insight are often overlooked, despite their importance in clinical settings where transparency and explanation of predictions are vital.

To address these gaps, this study makes several key contributions to the field. First, it provides a comprehensive comparison of widely used ML models—Naive Bayes, K-Nearest Neighbour, Support Vector Machine, Random Forest, C4.5, Logistic Regression, Linear Discriminant Analysis and Artificial Neural Network—on a common ATL-specific dataset (GSE33615). Second, beyond evaluating classification accuracy, the study applies model-specific feature importance techniques (Information Gain Ratio, SHAP etc.) extract informative gene signatures relevant to ATL. Third, the results reveal a consistent set of potential biomarkers (PLK3, ZSCAN18, and NELL2) identified across models, offering novel insight into the genomic basis of ATL and opening new possibilities for early diagnosis, targeted therapy, and drug development.

The sections of this paper are organized as follows: In the next section (Related Works), the works that have examined the use of machine learning in ATL diagnosis are summarized. The dataset used in the study is described in the Materials section. The details of the experimental design and the methods used are given in the Methods section. The performance comparisons of the models and the significant genes identified for each model are presented in the Results section. The Conclusion section includes contributions of the findings of this study and recommendations for future work.

2. Related Work

In recent years, ML has been increasingly adopted in biomedical sciences to support disease diagnosis, prognosis, and decision-making. Numerous studies demonstrate the diagnostic utility of ML in various domains, including hematologic malignancies, solid tumors, and even non-medical pattern recognition problems. However, the use of ML in ATL remains limited, particularly with regard to model interpretability and gene-level insights. Notable ML-based studies in this domain are critically reviewed below, with an emphasis on their methodological strengths, key findings, and limitations. Based on this review, key gaps in the existing literature were identified, which guided the motivation for this study. The novel contributions of the present work, along with its distinctions from prior research, are articulated in the final paragraph of this section.

Chong et al. (2020) employed a decision tree-based model for lymphoid neoplasms, achieving 94.7% overall accuracy. Nevertheless, the model failed to generalize well to ATL samples, showing error rates up to 100% in some cases—highlighting the diagnostic complexity of ATL using conventional ML tools.

Ghobadi et al. (2022) emphasized that although there are clinical guidelines for ATLlymphoma and its subtypes, they are far from being the gold standard. Therefore, reliable biomarkers should be found. With this aim, they proposed an SVM-based ML model capable of making diagnoses based on mRNA and miRNA features. Although the application details of the model were not given, it was emphasized nearly 95% accuracy was achieved. The related study is one of the rare studies in which disease-causing gene profiles are identified. However, these gene profiles were not extracted from the SVM model. Instead, experimentally validated target genes of miRNAs were examined.

In another study for T cell lymphoma diagnosis, deep learning was preferred (Xu et al., 2023). In the related study, class imbalance was observed in the dataset. To mitigate the potential negative impact of this issue on prediction performance, the authors used the bootstrap sampling to synthetically balance the class distribution. The classification performance of the model was calculated by area under curve (AUC) metric and found to be 0.75. Although the study highlights that the DL model detects gene expressions, the specific mechanism by which it accomplishes this remains unclear. Considering that the deep learning model is a black box approach, it is critical to explain in detail how these expressions are revealed from the model.

Another study using deep learning was conducted by Akalın and Yumuşak (2023). The dataset used in the study has 12500 gene profiles for each individual. In order to reduce the computational complexity, feature selection was performed using whale optimization. Long term short term memory (LSTM) model was used for the diagnosis of ALL, AML and MLL leukaemia types. The accuracy of the model was 89.88%. Since the method is structurally more suitable for time series analysis, its use may be limited in studies that do not examine time or memory dependent variation. Therefore, the suitability of the LSTM method in the context of this study should be evaluated. Regardless of the methodology, this study also did not examine the factors affecting the diagnosis.

Patel et al. (2021) proposed a multi-class diagnostic model for leukaemia diagnosis. Leukaemia tissues were micro-arrayed for extracting the gene expressions, then feature selection was applied on these expressions. Although the study was conducted for multi-class classification, logistic regression was executed based on a binary classification approach. Since the model was implemented using a "one-vs-all" strategy, the performance results are far from reflecting the true nature of the multi-class classification problem. Similar to many other studies in the domain, this work focuses solely on the performance of the model. It does not examine the disease-causing or protective gene profiles.

Study by Zhang et al. (2022), employed autoencoders to reduce the dimensionality of transcriptomic data, followed by clustering algorithms to identify biologically meaningful patient subgroups. The model successfully stratified patients into high- and low-risk categories, as evidenced by significant differences in survival analyses. Although the model identified molecular subtypes correlated with prognosis, its clinical applicability remains untested. Additionally, the use of black-box models like autoencoders limits interpretability, and no external validation was conducted to assess generalizability.

In broader contexts, although not directly related to the present topic, the examination of the following studies may also provide valuable insights.

A hybrid ensemble approach integrating logistic regression, support vector machines (SVM), and Extra Trees classifiers was implemented to enhance gene selection and prediction performance (Ruppapare et. al, 2022). The study utilized ADASYN for class imbalance and Chi-Squared tests for feature selection. Reported performance metrics were promising, with accuracy at 92%, F1-score at 90%, and balanced accuracy at 89%. However, the model's applicability to other hematological malignancies was not explored, and the biological significance of the selected genes was insufficiently discussed. External validation was also absent.

The work of Stagno et al. (2025) reflects the growing interest in applying ML to hematological malignancies. Their study reviews ML's utility in the diagnosis, prognosis, and treatment of chronic myeloid leukemia (CML), emphasizing the need for integrating predictive performance with clinical interpretability. Although insightful, this work focuses on CML and does not address ATL-specific genomic signatures.

Erdem and Bozkurt (2021) performed a comparative evaluation of various supervised ML techniques for prostate cancer prediction, showing that performance varies significantly across algorithms. However, similar to ATL studies, their focus remains largely on classifier performance rather than model explainability or biomarker identification.

To summarize the aforementioned studies, Table 1 provides an overview of key machine learning-based approaches applied to ATL and related leukemia diagnoses, outlining the datasets used, methodologies implemented, reported performance metrics, and notable limitations or contributions of each study.

Table 1. Comparison of the related work with the proposed study

Study (Author, Year)	Dataset Used	Methods Applied	Performance Metrics	Notes on Limitations/Strengths
Chong et al., 2020	Lymphoid neoplasm data	Decision Tree	Accuracy: 94.7% (but poor ATL generalization)	Failed to generalize to ATL; up to 100% error in ATL cases
Ghobadi et al., 2022	mRNA and miRNA data	SVM	Accuracy: ~95% (exact AUC not reported)	Gene profiles not derived from model; lack of application details
Xu et al., 2023	T-cell lymphoma gene expression	Deep Learning (DL) + Bootstrap for imbalance	AUC: 0.75	Black-box model; mechanism for gene expression detection unclear
Akalın & Yumuşak, 2023	12,500 gene profiles	LSTM + Whale Optimization	Accuracy: 89.88%	Suitable for time series; no diagnostic factor analysis
Patel et al., 2021	Leukemia microarray data	Logistic Regression (One-vs-All for multi-class)	Not specified (performance incomplete)	Lacks true multi-class representation and gene-level interpretation
Zhang et al., 2022	Transcriptomic data	Autoencoder + Clustering	Survival stratification	Lacks external validation; black-box limits interpretability
Ruppapare et al., 2022	Not ATL-specific (general genes)	Hybrid Ensemble (LR, SVM, Extra Trees) + ADASYN	Acc: 92%, F1: 90%, Bal. Acc: 89%	No biological interpretation; not ATL-specific
Stagno et al., 2025	CML-related	Literature Review of ML tools	—	Not ATL-specific; emphasizes need for interpretable ML
Erdem & Bozkurt, 2021	Prostate cancer data	Various supervised ML models	—	No gene-level interpretation; general ML benchmark

Proposed Work	GSE33615 (ATL-specific)	Naive Bayes, K-Nearest Neighbour, Support Vector Machine, Random Forest, C4.5, Logistic Regression, Linear Discriminant Analysis and Artificial Neural Network + SHAP & Info Gain + Evolutionary Opt.	AUC: 0.98 (ANN), F1: 0.95 (SVM), Precision: 0.95 (RF)	Model interpretability via SHAP, biologically relevant gene signatures identified, comparative model evaluation

Based on the literature review and the studies examined, despite the methodological progress in the field, three core gaps remain:

- Insufficient ATL-specific focus: Despite ATL's clinical importance, only a limited number of studies address it specifically, and even fewer examine gene-level signatures tied to its pathogenesis by using ML.
- Lack of comparative model analysis on the same dataset under controlled conditions for ATL.
- Single-method dependency: Most studies focus on a single classifier, without comparative benchmarking of different algorithms under the same experimental conditions
- Limited use of variable importance or explainable AI tools (e.g., SHAP, Information Gain) to extract disease-relevant gene signatures.

The proposed study addresses these shortcomings by:

- Comparatively evaluating multiple ML models (SVM, RF, C4.5) under identical conditions on the GSE33615 dataset
- Combining high prediction performance with biological interpretability, offering a framework for biomarker discovery, therapeutic target identification, and improved clinical decision-making in ATL.
- Applying model-specific variable importance techniques (SHAP for SVM and RF, Information Gain Ratio for C4.5) to identify biologically meaningful genes associated with ATL.

In contrast to prior literature that prioritizes accuracy alone, this study not only advances diagnostic performance but also contributes to understanding the molecular basis of ATL by integrating prediction with gene-level interpretation. Thus setting a foundation for interpretable precision oncology in the context of AT and paving the way for biomarker-driven research and targeted therapeutic strategies

3. Materials and Methods

This study was conducted with two main motivations. The first one is to compare the performance of ML methods that are widely used in T-cell leukemia diagnosis and to establish a final model with high diagnostic performance. The second motivation is to identify disease-causing genes for early diagnosis of ATL. The general infrastructure of the study is presented in Figure 1.

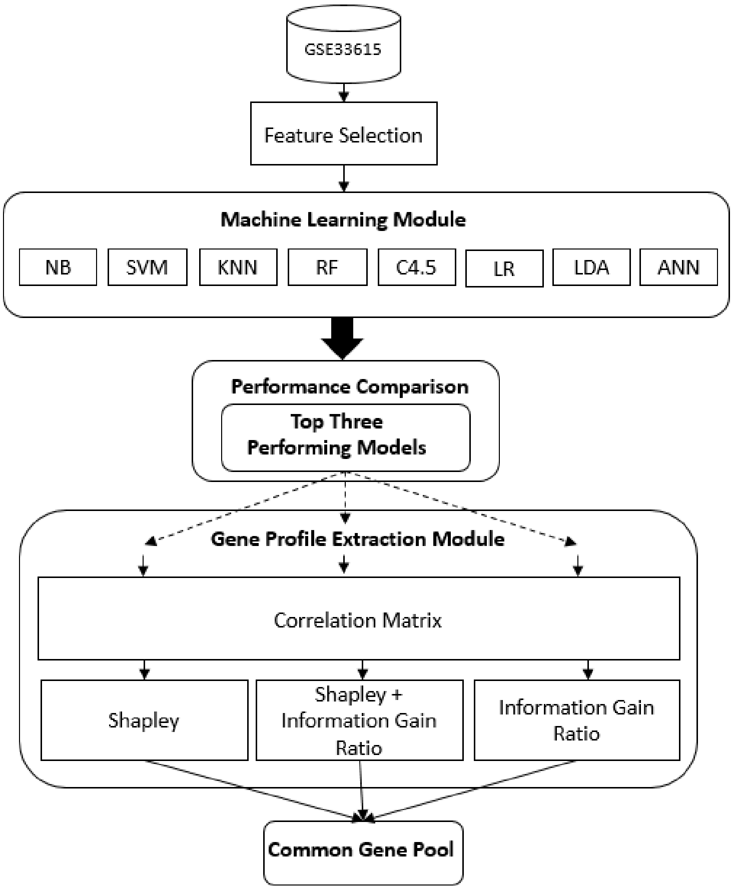

Figure 1. Infrastructure of the experimental design

In this study, the leukemia dataset, GSE33615, was retrieved from Gene Expression Omnibus (GEO) database. Genes with low variance were eliminated in the feature selection step. Then, several ML models were trained and validated. The performance of each model was compared according to precision, recall, AUC and F1 score metrics. In order to find the genes affecting ATL, the results of top tree performing models were first given to the correlation matrix. To calculate the variable importance for each ML model, different metrics were applied. Shapley's criterion and variants were used for black-box based methods. TreeSHAP and information gain ratio (IGR) was used for RF.

Then, the significant genes for each model were ranked and the common ones were added to the gene pool as genes affecting the disease.

3.1. Material

To identify marker genes in leukemia cancer, the dataset GSE33615 (Fujikawa et al., 2016; Yamagishi et al., 2012) from the Gene Expression Omnibus (GEO) database was used. The dataset consists of 52 ATL diagnosed cases and 21 controls, each containing 45,015 gene profiles. For the preliminary analysis of these data, the box plot, Figure 2, was used.

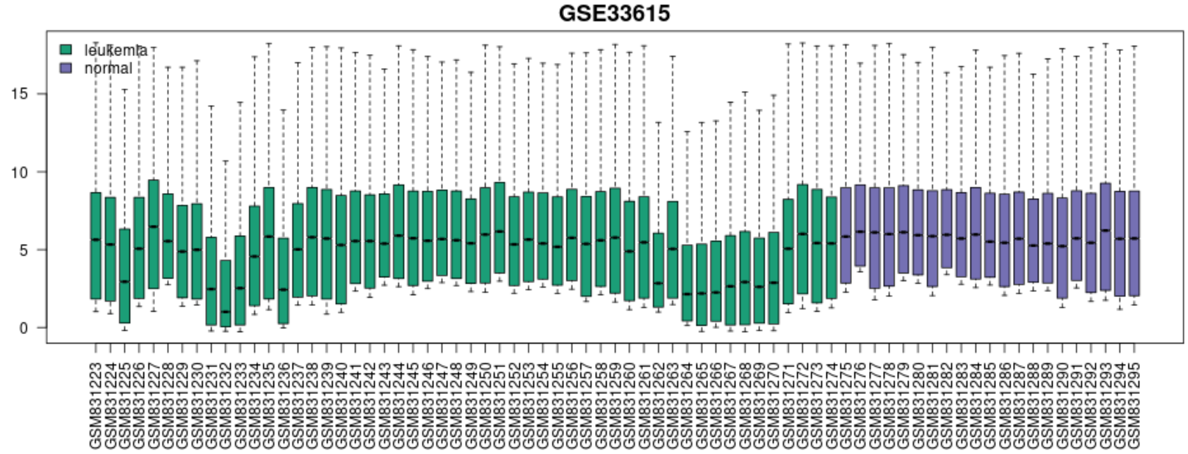

Figure 2. Distribution of the GSE33615 Data set. Purple color indicates CD4+t cell (control), Green Color indicates ATL (case).

Both groups have a wide range of genes; however, they exhibit significant expression differences. Figure 2 shows that leukemia samples generally have higher gene expression levels and a wider distribution range than normal samples. This indicates that the dataset has a wide range of gene expressions and that gene markers in the presence of ATL are different from normal samples.

3.2. Methods

The performance of DT based methods, specifically C4.5 and RF and black-box approaches such as SVM, have been compared for the diagnosis of ATL. Unlike other studies in the literature, this study aims to identify the gene profiles affecting the disease along with high diagnostic performance. Therefore, a specific variable importance extraction was performed for each model, and the common genes identified as significant in the diagnosis of ATL were determined. The methodology used to achieve these primary objectives are presented in detail under subheadings. All ML and feature selection algorithms were developed in the R programming language.

3.2.1. Feature Selection

The GEO GSE33615 dataset was used for training and validation of the ML models. This set consists of 52 cases and 21 controls with 45,015 gene profiles for each individual. A preprocessing step was applied in order to improve the performance of the models and find the genes with the highest impact on diagnosis.

In this step, genes exhibiting low genetic diversity were filtered out based on their variance values. Specifically, the varFilter and genefilter functions from the R platform were employed with a threshold value of 0.9. Utilizing such a high variance threshold is a recognized strategy to enhance the reliability of data analysis by focusing on genes with significant expression variability. This approach effectively reduces noise and potential false positives, thereby improving the accuracy of the model. For instance, Haury et al. (2011) demonstrated that applying stringent variance thresholds can lead to more stable and interpretable molecular signatures in high-dimensional gene expression data. Similarly, Lee et al. (2013) highlighted the importance of robust feature selection methods in early cancer detection, emphasizing that higher variance thresholds contribute to the identification of consistent and biologically relevant biomarker

3.2.2. Naive Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem, assuming feature independence given the class label (Rish, 2001). It is particularly effective for high-dimensional data and works efficiently even with small datasets.

The posterior probability that a given instance $x=(x_1,,x_2,...,x_n)$ belongs to class C_k is computed as in Equation 1:

$$P(C_k|x) = \frac{P(C_k)\prod_{i=1}^{n} P(C_k|x_i))}{P(x)} \tag{1}$$

Here, $P(C_k)$ is the prior probability of class C_k, $P(x_i/C_k)$ is the likelihood of feature x_i given class C_k and $P(x)$ is the evidence.

3.2.3. K Nearest Neighbours

KNN is a simple yet effective non-parametric classification algorithm that makes decisions based on proximity between data points in a feature space (Cover and Hart, 1967). The method does not construct an explicit model during training; instead, it stores the entire training dataset and performs classification during the prediction phase based on distance calculations (Keller et al., 1985).

In this algorithm, a given instance is assigned to the class most common among its k nearest neighbors, where k is a predefined positive integer. The proximity between data points is usually measured using distance metrics such as Euclidean distance, Manhattan distance, or Minkowski distance. In this study, Euclidean distance is used, which is calculated as in Equation 2 for a d-dimensional feature space:

$$D(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \quad (2)$$

Here, *x* and *y* are the feature vectors of two instances, and *d* is the total number of features. Once the distances to all training instances are computed, the *k* closest instances are selected, and the class label is determined by majority voting. In the case of a tie, methods such as distance-weighted voting can be applied.

3.2.4. Support Vector Machine

SVM, developed by Vapnik et al. (1995), is one of the most preferred ML methods in the literature due to its successful performance on high-dimensional datasets of nonlinear problems (Roy and Chakraborty, 2023). Its main goal is to find the hyperplane that will provide the highest margin between classes.

When the classes are linearly separable, a dataset of *n* points can be represented as *(x_1,y_1),..., (x_n,y_n)*. Accordingly, any hyperplane, where w^T is the weight vector, *x* is the input vector and *b* is the bias, can be written as the set of points given in Equation 3:

$$w^T x + b = \quad (3)$$

In binary classification, this hyperplane separates the classes according to the following equation:

$$f(x) = \{1, w^T - b \geq 1 \ -1, w^T x + b \leq -1 \quad (4)$$

The SVM tries to optimize the distance between these two hyperplanes to the maximum margin by minimizing $||w||$.

When the classification problem is nonlinear, the data is mapped to a higher dimensional space using the kernel method instead of dot product. Different kernel functions (Equations 5 to 8) can be used for this purpose.

$$K(x_i, x_j) = x_i x_j \quad (5)$$

$$K(x_i, x_j) = (x_i x_j + c)^d \quad (6)$$

$$K(x_i, x_j) = exp\ (-\gamma||x_i - x_j||^2) \quad (7)$$

$$K(x_i ,x_j) = tanh\ (\alpha x_i x_j + c) \qquad (8)$$

In this study, the Radial-based SVM, renowned for its ability to model complex relationships in biological data, was employed due to its effectiveness in high-dimensional and small datasets, yielding significant results in genetics and molecular biology (Guido et al., 2024).

3.2.5. Random Forest

RF is an ensemble model consisting of multiple DTs (Breiman, 2001). Each tree is trained using a random subset and the final decision is determined by the majority voting. RO increases the generalization ability of the model by reducing overfitting (Breiman, 2001).

N denotes the total number of trees, each tree $h_i(x)$ predicts the data x and the results are combined with the ensemble model as in Equation 9.

$$f(x) = \frac{1}{N}\sum_{i=}^{N} h_i(x) \qquad (9)$$

3.2.6. C4.5 Decision Tree

C4.5 (Quinlan, 2014) is an advanced version of the ID3 algorithm and stands out with its ability to work with both categorical and numerical data (Fahim et al., 2023). The model selects the attributes that will divide the dataset in the most homogeneous way possible and places them in nodes at different levels in the tree structure. Different criteria can be used for this division (branching).

In this study, the information gain ratio (IGR) based on entropy calculation is chosen as the branching criterion. Entropy is calculated as in Equation 10. where S is the dataset and p_i is the probability for class i.

$$H(S) = -\sum_{i=1}^{n} p_i(p_i) \qquad (10)$$

After the entropy is calculated, the IGR for each variable is calculated as in Equation 11.

$$IG(T,X) = H(T) - \sum_{V \in Val(X)} \frac{|T_v|}{T} H(T_v) \qquad (11)$$

T is a dataset and X is a variable and T_v is a subset of X with value in Equation 2. The variable with the highest IGR is assigned to the relevant node of the tree.

3.2.7. Logistic Regression

Logistic Regression is a widely used linear classification algorithm that models the probability of a binary outcome using the logistic function (Hosmer et al., 2013). It estimates the parameters by maximizing the likelihood function based on the observed class labels.

For a given feature vector *x*, the probability that the instance belongs to class 1 is given by:

$$P(y=1|x)=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \quad (12)$$

Where β_0 is the intercept, and $\beta_1,\ldots,\beta_n$ are the model coefficients.

3.2.8. Linear Discriminant Analysis

Linear Discriminant Analysis is a supervised classification method that assumes a Gaussian distribution for each class and models a common covariance matrix across classes (Fisher, 1936). It seeks a linear combination of features that best separates two or more classes.

In case of binary classification, the decision function is based on the linear discriminant score given in Equation – 13:

$$\delta_k(x)=x^T\,\Sigma^{-1}\mu_k-{}^1\!/_2\,\mu_k{}^T\,\Sigma^{-1}\mu_k+\log\pi_k \quad (13)$$

μ_k is the mean vector of class *k*, Σ is the shared covariance matrix, π_k is the prior probability of class *k*.

An observation is assigned to the class with the maximum discriminant score. In this study, LDA was chosen due to its efficiency and robustness with small datasets and linearly separable features.

3.2.9. Artificial Neural Networks

Artificial Neural Networks are inspired by the biological structure of the human brain and consist of interconnected nodes (neurons) arranged in layers: input, hidden, and output (Haykin, 1999). Each neuron computes a weighted sum of its inputs and passes it through a nonlinear activation function. For a single hidden layer, the output of the network can be expressed as:

$$y=f(\textstyle\sum_{j=1}^{h} w_j^{(2)}\,.\,\sigma\left(\sum_{i=1}^{n} w_{ij}^{(1)}x_i+b_j\right)+b_0) \quad (14)$$

x_i are the input features, $w_{ij}{}^1$ and $w_j{}^2$ are the weights for input-to-hidden and hidden-to-output layers respectively, b_j and b_0 are bias terms, $\sigma(\cdot)$ is the activation function (e.g., ReLU, sigmoid) and $f(\cdot)$ denotes the activation function in the output layer, typically softmax or sigmoid depending on the classification type

All methods used in the study are implemented in R language and the hyper-parameter values were optimized by evolutionary algorithm (Lee et al., 2021).

3.2.10. Gene Profiling Module

For top tree performing ML models, this module ranks the genes affecting the classification according to their importance. For this ranking, model-specific variable importance method is applied. For the tree-based models IGR and TreeSHAP (Inan and Rahman, 2023; Lundberg et al., 2019) was used. For the ranking of the attributes in SVM and ANN model, the SAHPLEY criterion (S. Lundberg, 2017), and KernelSHAP (Ekanayake et al., 2022) which are widely used in black box approaches, is preferred. Based on game theory, this criterion calculates the presence or absence of a selected attribute in the model using different combinations of attributes. The SHAPLEY measure of the selected attribute is calculated by weighting the combinations based on the values of the attribute as in Equation 15.

$$\emptyset_{j(val)} = \sum_{S\epsilon\{1,\dots,p\}\backslash j} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{j\} - val(S)) \tag{15}$$

In the training set consisting of p features, let S be a subset of features, x represent a vector containing the values of the selected feature, and $val_x(S)$ denote the prediction of the values of features in S based on the values of features not included in S.

4. Findings and Discussion

One of the motivations of the study is to compare the performance of ML methods frequently used for the diagnosis of ATL. For this purpose, NB, KNN, SVM, RF, C4.5, LR, LDA and ANN models were built, trained and tested on the Gene Expression Omnibus (GEO) GSE33615 dataset. A 5-fold cross-validation was applied for the training and testing process. For each fold, random stratified sampling was used.

To assess potential overfitting, the difference in training and validation performance was evaluated based on AUC. The corresponding comparison is presented in Figure 3.

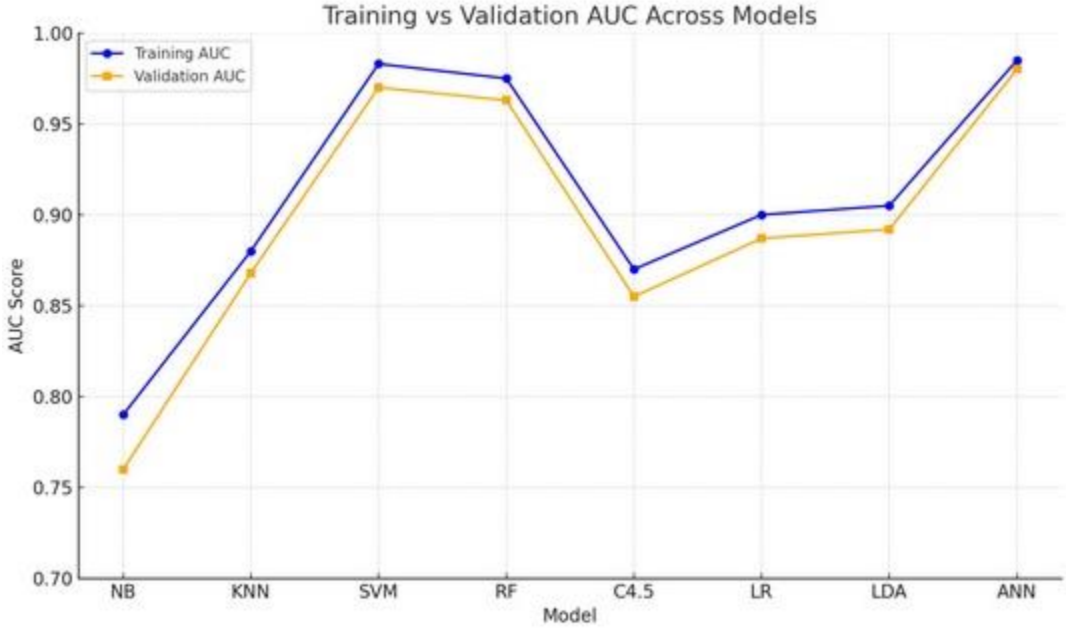

Figure 3. Train and validation AUC comparison of the esablished models.

Figure 3 demonstrates that the training AUC and Validation AUC exhibit a similar trend across all models. Moreover, there is no significant performance gap between training and validation. This suggests that the models learned without overfitting. Significant overfitting typically occurs when the training AUC is substantially higher than the validation AUC (Goodfellow et al., 2016). Therefore, a low but consistent validation AUC indicates that the model generalizes effectively.

Following the overfitting analysis, the validation performance of the models was compared using precision, recall, AUC and F1 metrics. The results of this comparison are presented in Table 2.

Table 2. Performance comparison of the models

MODEL	Precisio n	Recall	AUC	F1
NB	0.576	0.852	0.760	0.688
KNN	0.81	0.95	0.868	0.874
SVM	0.92	0.98	0.970	0.949
RF	0.95	0.945	0.963	0.947
C4.5	0.841	0.835	0.855	0.838
LR	0.840	0.835	0.887	0.837
LDA	0.872	0.823	0.892	0.846
ANN	0.928	0932	0.980	0.930

Table 2 shows that ANN, SVM and RF outperform other models in terms of AUC and F1. ANN model has the highest AUC, while SVM has the highest F1. The superiority of SVM and ANN over other models can be explained by their high ability to adapt to nonlinear problems on multidimensional data (Vapnik, 1995; LeCun et al., 2015). On the other hand, the strong performance of the RF model can be attributed to its ensemble learning structure, which effectively reduces overfitting and handles high-dimensional, nonlinear data (Breiman, 2001; Chi et al., 2022).

The main motivation and the contribution of the study is the extraction of gene profiles causing ATL from the ML methods. For this purpose, genes that are highly correlated with ATL positive status were selected from top tree performing models (ANN, SVM, RF). Then, different variable importance calculations were made on each model. IGR and TreeSHAP (Inan and Rahman, 2023) were used for RF, SHAPLEY (Wang et al., 2024) and KernelSHAP (Ekanayake et al., 2022) was used for ANN and SVM. As a result of these calculations, the top ten genes found significant for each model are presented through tables 3-5.

Table 3. Top ten significant genes found by ANN model. The genes found to be significant by all models are shown in bold italics.

	SYMBOL	**GENE NAME**
1	ARHGAP20	Rho GTPase Activating Protein 20
2	FMNL2	Formin Like 2
3	***PLK3***	Polo-like kinase 3
4	SPHK2	Sphingosine Kinase 2
5	ZBTB40	Zinc Finger And BTB Domain Containing 40
6	***ZSCAN18***	Zinc finger and SCAN domain-containing 18
7	CDHR1	Cadherin Related Family Member 1
8	PRPSAP2	Phosphoribosyl Pyrophosphate Synthetase Associated Protein 2
9	TIAM2	TIAM Rac1 Associated GEF 2
10	STAT4	Signal Transducer and Activator of Transcription 4

The first three genes found to be important by the ANN model are known to be closely associated with leukemia ATL and cancer mechanisms. The ARHGAP20 gene encodes a Rho GTPase-activating protein that regulates cytoskeletal dynamics and cell motility. In chronic lymphocytic leukemia (CLL), ARHGAP20 expression was unexpectedly higher in cases (Liu et al., 2021). FMNL2 is an actin nucleating protein that promotes cell migration and invasion. While specific studies on FMNL2 in leukemia are limited, its role in cytoskeletal regulation implicates it in various cancers (Zhu et al., 2011). PLK3 is a serine/threonine kinase involved in cell cycle regulation and DNA damage response. Although direct studies linking PLK3 to leukemia are scarce, its function in cell cycle control suggests potential involvement in hematologic malignancies (Zhang et al., 2021;) Hukasova, 2017; Xie et al., 2001)

The top ten significant genes found by RF model is presented in Table 4.

Table 4. Top ten significant genes found by RF model. The genes found to be significant by all models are shown in bold italics.

	SYMBOL	GENE NAME
1	***ZSCAN18***	Zinc finger and SCAN domain-containing 18
2	STAT4	Signal Transducer and Activator of Transcription 4
3	ITK	IL2-inducible T-cell kinase
4	***PLK3***	Polo-like kinase 3
5	CDKN1A	Cyclin-dependent kinase inhibitor 1A
6	DEFA4	Alpha-defensing 4
7	BCL11B	B-cell chronic lymphocytic leukemia/lymphoma 11B
8	RGS16	G-protein signaling regulator 16
9	***NELL2***	Neural EGFL like 2
10	PGRMC2	Progesterone receptor membrane component 2

Top three genes in Table 4, ZSCAN18, STAT4 and ITK, are associated with leukemia in the literature. ZSCAN18 is a member of the zinc finger protein family that regulates gene expression and therefore plays a role in hematopoiesis (Hall, 2021). STAT4 is involved in many cytokine and growth factor signaling pathways that may affect leukemia cell survival, proliferation and apoptosis (Frank, 1999; Rajasingh et al., 2006). It is also involved in immune response regulation. Recent studies have identified STAT4 as a prognostic biomarker in AML, where its expression correlates with disease progression (Li et al., 2024). lITK is an important tyrosine kinase in T-cell receptor signaling and is associated with T-cell acute lymphoblastic leukemia (T-ALL) (Cordo et al., 2022). Aberrant ITK activity has been implicated in T-cell malignancies, including ATL, by promoting uncontrolled T-cell proliferation and survival Abnormal activation or mutations of these genes may be involved in the pathogenesis of the disease. Therefore, it may be useful to target these features of genes to propose new therapeutic strategies.

The top ten significant genes found by SVM model is presented in Table 5.

Table 5. Top ten significant genes found by SVM model. The genes found to be significant by all models are shown in bold italics.

	SYMBOL	GENE NAME
1	CCR7	C-C motif chemokine receptor 7
2	HIP1R	Huntingtin interacting protein 1 related
3	LYSMD2	LysM domain containing 2
4	***PLK3***	Polo-like kinase 3
5	***ZSCAN18***	Zinc finger and SCAN domain-containing 18
6	***NELL2***	Neural EGFL like 2
7	CXorf57	Chromosome X open reading frame 57
8	ZNF502	Çinko parmak protein 502
9	ITPKB	Inositol-trisphosphate 3-kinase B
10	CDKN1A	Cyclin-dependent kinase inhibitor 1A

The first three genes found significant by SVM were CCR7, HIP1R and LYSMD2, respectively. CCR7, which plays an important role in the spread of leukemia cells, is a chemokine receptor in lymphocyte migration and homing processes (Choi et al., 2020; Legler et al., 2014). CCR7 is frequently expressed in ATL cells, facilitating their migration to lymphoid tissues through interactions with its ligands, CCL19 and CCL21. Gain-of-function mutations in CCR7 have been observed in ATL patients, leading to enhanced downstream signaling and potentially contributing to disease progression (Sakamoto et al., 2022).HIP1R is associated with cytoskeleton and vesicular traffic and may be involved in cell growth and cancer development (Hyun and Ross, 2004; Saralamma et al., 2020). LYSMD2 is a LysM domain-containing protein and is involved in immune responses and cellular processes (Miao et al., 2023; Sundaramurthi et al., 2023).

The intersection set of genes found to be significant by all three models showed that PLK3, ZSCAN18 and NELL2 genes were prominent in ATL diagnosis. These genes are implicated in various malignancies, including cancer, leukemia, and ATL.

Among these genes, PLK3 is a cell cycle regulator and plays a role in DNA damage response processes (Helmke et al., 2016; Hukasova, 2017) . It is a serine/threonine kinase involved in cell cycle regulation and stress response. Unlike its family members PLK1 and PLK2, which are often overexpressed in cancers, PLK3 acts as a tumor suppressor (Goroshchuk et al., 2019). It mediates apoptosis and responds to DNA damage and oxidative stress. Aberrant expression of PLK3 has been observed in various tumors, suggesting its role in tumorigenesis. In acute leukemia, PLK3's tumor-suppressive function underscores its potential as a therapeutic target (Helmke et al., 2016).

The effects of different protein isoforms of the ZSCAN family on cancer cell growth and proliferation are being investigated. It is a validated prognostic marker in renal cell carcinoma (KIRC), and breast cancer where higher expression correlates with favorable outcomes (Wang et al., 2023). Therefore, it can be considered as an important cancer marker (Li et al., 2023). However, its role in leukemia and ATL remains unclear due to limited data.

NELL2 is a secreted glycoprotein involved in neural development and chromatin remodeling. In Ewing sarcoma, NELL2 autocrine signaling enhances cell proliferation by inhibiting cdc42 and promoting BAF complex assembly (Nakamura et al., 2015). This signaling pathway also upregulates EWS-FLI1 transcriptional output. Although NELL2 is not a prognostic marker in glioblastoma, its expression is elevated in thyroid carcinoma, suggesting its involvement in certain cancers (Jayabal et al., 2021).

Beyond the genes commonly identified across all models, several others—including BCL11B, CDKN1A, DEFA4, RGS16, SPHK2, —were also found to be associated with leukemia and ATL, underscoring their potential relevance in disease pathogenesis.

BCL11B is a transcription factor essential for T-cell development, it exhibits reduced expression in ATL cases. Studies have shown that ectopic expression of BCL11B suppresses the growth of ATL-derived cell lines, suggesting its role as a tumor suppressor in ATL pathogenesis (Kurosawa et al., 2013).

CDKN1A is typically overexpressed in HTLV-1-infected cell lines and its expression is downregulated in primary ATL cells (Watanabe et al., 201). This downregulation is often due to promoter methylation, implicating epigenetic modifications in ATL development (Cordo et al., 2022).

DEFA4 encodes an antimicrobial peptide predominantly expressed in neutrophils. Elevated DEFA4 expression has been reported in acute myeloid leukemia (AML) patients, suggesting its potential as a biomarker for disease progression (Zhao et al., 2023).

RGS16 modulates G-protein-coupled receptor signaling pathways. Altered RGS16 expression has been associated with various hematological malignancies, including leukemia, indicating its role in leukemogenesis (LeBlanc et al., 2020).

SPHK2 is involved in sphingolipid metabolism and has been found to be overexpressed in large granular lymphocyte leukemia. SPHK2 promotes cell survival through upregulation of anti-apoptotic proteins like Mcl-1, highlighting its potential as a therapeutic target (LeBlanc et al., 2020).

A deeper study of these genes may help better understand the mechanisms of the disease and their genetic interactions. The results of this study represent an important step forward in understanding the complexity of ATL and identifying potential therapeutic targets through detailed analysis of genetic data.

Despite the promising diagnostic performance of the proposed models and the identification of meaningful gene signatures, this study has several limitations. Firstly, the methodology was evaluated using a single publicly available dataset (GSE33615), which may not fully capture the diversity of ATL cases across different populations. Secondly, although interpretability was enhanced through model-specific variable importance measures (e.g., SHAP, Information Gain Ratio), these methods do not entirely explain complex gene-gene interactions or account for biological noise in gene expression data. Additionally, the models assume the availability and accuracy of all relevant features, which may not always be feasible in real-world clinical datasets. These limitations suggest that external validation and further biological investigation are essential for confirming the robustness and clinical applicability of the findings

5. Conclusions and Recommendations

ATL is a highly aggressive disease caused by human T-cell leukemia virus type I (HTLV-1) with an extremely unfavorable prognosis. It is usually diagnosed after symptoms become apparent, which can make it difficult to control the disease and provide effective treatment. It is therefore critical to find models and biomarkers that can be used for early diagnosis of the disease.

There are different studies using ML for early diagnosis of the related disease. When these studies are analyzed, it is observed that the most preferred methods are tree-based methods such as DT and RF and black box-based methods such as SVM. Some of these studies show that ATL derivatives are very difficult to diagnose (Chong et al., 2020). For this reason, current studies have used a single classifier model and the diagnostic performance has been the focus of the study. The underlying reasons for the model's diagnostic performance are not analyzed.

In this study, the diagnostic performances of several machine learning methods (Naive Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, C4.5, Logistic Regression, Linear Discriminant Analysis and Artificial Neural Network), which are frequently used in the literature for ATL diagnosis, were first compared according to precision, recall, AUC and F1 score metrics on the Gene Expression Omnibus GSE33615 dataset. Among these, Artificial Neural Network stood out with an AUC of 0.98 and an F1 score of 0.93. It was followed by SVM with an AUC of 0.97 and 0.957 F1 score.

To fill the gap in the literature, the main contribution of this study is to reveal the gene profiles that have an effect on ATL diagnosis. For this purpose, the ten most significant gene profiles affecting the diagnosis were extracted from the top tree performing models (ANN, SVM and RF) by using different variable importance calculations. Examining the intersection set of these genes, PLK3, ZSCAN18 and NELL2 stood out in distinguishing the ATL positive class. It is seen that both the individual gene clusters of the models (Table 3, Table 4, and Table 5) and the genes in the intersection set are in accordance with cancer and leukemia studies in the literature. Among these genes, PLK3 is a cell cycle regulator and plays a role in DNA damage response processes. ZSCAN18 is a transcription factor containing zinc finger and SCAN domains that may play a role in transcriptional regulation and cell differentiation. This protein regulates gene expression by binding to DNA and shows RNA polymerase II-specific transcription factor activity (Wang et al., 2023). NELL2 is the gene encoding a protein that plays an important role in nervous system development and synaptic plasticity. This protein is active in cell signaling and neuronal differentiation processes (Nakamura et al., 2015).Therefore, PLK3, ZSCAN18, NELL2 can be considered as a significant cancer marker. Beyond the genes commonly identified across all models, several others—including BCL11B, CDKN1A, DEFA4, RGS16, SPHK2, —were also found to be associated with leukemia and ATL, underscoring their potential relevance in disease pathogenesis.

The effects of the identified genes on cancer biology and their contribution to potential treatment strategies are highly valuable. A detailed understanding of the effects of these genes on cancer development and responses to treatment may enable the development of more effective and personalized treatment methods. It is important to evaluate the effects of these genes on cancer development, metastasis processes and responses to treatment in more depth. For this reason, it may be recommended to examine the functions and interactions of these genes in more detail in future studies.

6. Future Work

For future research, expanding the methodological framework to include ensemble and deep learning-based models could further improve diagnostic accuracy and reveal more complex patterns in gene expression. Although deep learning approaches, such as convolutional and recurrent neural networks, offer significant predictive capabilities, their lack of transparency limits clinical usability. Therefore, integrating explainable AI techniques into such models (e.g., attention mechanisms, integrated gradients) may help address this issue. Additionally, hybrid ensemble strategies that combine the strengths of interpretable and high-performing models should be explored to achieve both accuracy and transparency in ATL diagnostics. This direction may contribute to developing more generalizable, reliable, and clinically relevant decision support tools.

Moreover, future studies should aim to increase the availability of datasets specific to ATL in order to evaluate model performance across diverse and unseen data. In cases where new datasets cannot be obtained, data augmentation techniques may be employed to expand the sample size and enhance the robustness and generalizability of the models

Authors' Contributions

All authors contributed equally to the study.

Statement of Conflicts of Interest

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The authors declare that this study complies with Research and Publication Ethics.

References

Abass, Y. A., & Adeshina, S. A. (2021). Deep learning methodologies for genomic data prediction. Journal of Artificial Intelligence for Medical Sciences, 2(1), 1-11

Akalın, F., and Yumuşak, N. (2023). Mikrodizi veri kümesindeki ALL, AML ve MLL lösemi türlerine ilişkin gen anomalilerinin LSTM sinir ağı ile sınıflandırılması. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 38(3), 1299–1306.

Breiman, L. (2001). Random Forests. *Mach Learn,* 45 (1): 5–32.

Chi, C. M., Vossler, P., Fan, Y., & Lv, J. (2022). Asymptotic properties of high-dimensional random forests. The Annals of Statistics, 50(6), 3415-3438.

Choi, H., Song, H., and Jung, Y. W. (2020). The roles of CCR7 for the homing of memory CD8+ T cells into their survival niches. *Immune Network*, 20(3).

Chong, Y., Lee, J. Y., Kim, Y., Choi, J., Yu, H., Park, G., Cho, M. Y., and Thakur, N. (2020). A machine-learning expert-supporting system for diagnosis prediction of lymphoid neoplasms using a probabilistic decision-tree algorithm and immunohistochemistry profile database. *Journal of Pathology and Translational Medicine*, 54(6), 462–470.

Cook, L., Rowan, A., & Bangham, C. (2021). ATLleukemia/lymphoma—Pathobiology and implications for modern clinical management. *Annals of Lymphoma*, 5.

Cordo, V., Meijer, M. T., Hagelaar, R., de Goeij-de Haas, R. R., Poort, V. M., Henneman, A. A., Piersma, S. R., Pham, T. V., Oshima, K., and Ferrando, A. A. (2022). Phosphoproteomic profiling of T cell acute lymphoblastic leukemia reveals targetable kinases and combination treatment strategies. *Nature Communications*, 13(1), 1048.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

Eckardt, J. N., Bornhäuser, M., Wendt, K., and Middeke, J. M. (2020). Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects. *Blood Advances*, 4(23), 6077-6085.

Ekanayake, I. U., Meddage, D. P. P., and Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials*, 16, e01059.

Erdem, E., & Bozkurt, F. (2021). A comparison of various supervised machine learning techniques for prostate cancer prediction. Avrupa Bilim ve Teknoloji Dergisi, (21), 610-620.

Fahim, N. I., Utsha, M. A. H., Karmaker, R. S., Ullah, M. O., and Farid, D. M. (2023). Decision Tree using Feature Grouping. *2023 26th International Conference on Computer and Information Technology (ICCIT)* (pp. 1–5). Cox's Bazar, Bangladesh.

Faiz, M., Mounika, B. G., Akbar, M., and Srivastava, S. (2024). Deep and Machine Learning for Acute Lymphoblastic Leukemia Diagnosis: A Comprehensive Review. *Advances in Distributed Computing and Artificial Intelligence Journal*, 13, e31420-e31420.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2), 179-188.

Frank, D. A. (1999). STAT Signaling in the Pathogenesis and Treatment of Cancer. *Molecular Medicine*, 5(7), 432–456. https://doi.org/10.1007/BF03403538

Fujikawa, D., Nakagawa, S., Hori, M., Kurokawa, N., Soejima, A., Nakano, K., Yamochi, T., Nakashima, M., Kobayashi, S., and Tanaka, Y. (2016). Polycomb-dependent epigenetic landscape in ATLleukemia. *Blood, The Journal of the American Society of Hematology*, 127(14), 1790–1802.

Gessain, A., and Cassar, O. (2012). Epidemiological aspects and world distribution of HTLV-1 infection. *Frontiers in Microbiology*, 3, 388.

Ghobadi, M. Z., Emamzadeh, R., and Afsaneh, E. (2022). Exploration of mRNAs and miRNA classifiers for various ATLL cancer subtypes using machine learning. *BMC Cancer*, 22(1), 433. https://doi.org/10.1186/s12885-022-09540-1

Goroshchuk, O., Kolosenko, I., Vidarsdottir, L., Azimi, A., & Palm-Apergi, C. (2019). Polo-like kinases and acute leukemia. Oncogene, 38(1), 1-16.)

Guido, R., Ferrisi, S., Lofaro, D., and Conforti, D. (2024). An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information*, 15(4), 235.

Guo, W., Liu, R., Ono, Y., Ma, A.-H., Martinez, A., Sanchez, E., Wang, Y., Huang, W., Mazloom, A., and Li, J. (2012). Molecular characteristics of CTA056, a novel interleukin-2-inducible T-cell kinase inhibitor that selectively targets malignant T cells and modulates oncomirs. *Molecular Pharmacology*, 82(5), 938–947.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.

Hall, E. S. (2021). *Applying Polygenic Models to Disentangle Genotype-Phenotype Associations across Common Human Diseases.* (unpublished master'sdissertation). University of Toronto, Canada

Haury, A. C., Gestraud, P., & Vert, J. P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PloS one, 6(12), e28210.

Haykin, S. (1994). Neural networks: a comprehensive foundation. Prentice Hall PTR

Helmke, C., Becker, S., and Strebhardt, K. (2016). The role of Plk3 in oncogenesis. *Oncogene*, 35(2), 135–147.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

Hukasova, E. (2017). *Cell Cycle Regulation and DNA Damage Response: A Record of Polo-Like Kinase 1 Activity*.(unpublished doctoral dissertation). Karolinska Institutet, Stockholm, Sweden

Hyun, T. S., and Ross, T. S. (2004). HIP1: Trafficking roles and regulation of tumorigenesis. *Trends in Molecular Medicine*, 10(4), 194–199.

Inan, M. S. K., and Rahman, I. (2023). Explainable AI integrated feature selection for landslide susceptibility mapping using TreeSHAP. *SN Computer Science*, 4(5), 482.

Ishitsuka, K., and Tamura, K. (2014). Human T-cell leukaemia virus type I and ATLleukaemia-lymphoma. *The Lancet Oncology*, 15(11), e517–e526.

Jayabal, P., Zhou, F., Lei, X., Ma, X., Blackman, B., Weintraub, S. T., ... & Shiio, Y. (2021). NELL2-cdc42 signaling regulates BAF complexes and Ewing sarcoma cell growth. Cell reports, 36(1)

Katsuya, H., Ishitsuka, K., Utsunomiya, A., Hanada, S., Eto, T., Moriuchi, Y., Saburi, Y., Miyahara, M., Sueoka, E., and Uike, N. (2015). Treatment and survival among 1594 patients with ATL. *Blood, The Journal of the American Society of Hematology*, 126(24), 2570–2577.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. IEEE transactions on systems, man, and cybernetics, (4), 580-585.

Kılıçarslan, S., and Pacal, I. (2023). Domates Yapraklarında Hastalık Tespiti İçin Transfer Öğrenme Metotlarının Kullanılması. *Mühendislik Bilimleri ve Araştırmaları Dergisi*, 5(2), 215–222.

Kim, D. H., Roh, Y.-G., Lee, H. H., Lee, S.-Y., Kim, S. I., Lee, B. J., and Leem, S.-H. (2013). The *E2F1* Oncogene Transcriptionally Regulates *NELL2* in Cancer Cells. *DNA and Cell Biology*, 32(9), 517–523. https://doi.org/10.1089/dna.2013.1974

Kurosawa, N., Fujimoto, R., Ozawa, T., Itoyama, T., Sadamori, N., & Isobe, M. (2013). Reduced level of the BCL11B protein is associated with adult T-cell leukemia/lymphoma. PLoS One, 8(1), e55147.)

LeBlanc, F. R., Pearson, J. M., Tan, S. F., Cheon, H., Xing, J. C., Dunton, W., ... & Loughran Jr, T. P. (2020). Sphingosine kinase-2 is overexpressed in large granular lymphocyte leukaemia and promotes survival through Mcl-1. British journal of haematology, 190(3), 405-417.) .

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Lee, H. W., Lawton, C., Na, Y. J., & Yoon, S. (2013). Robustness of chemometrics-based feature selection methods in early cancer detection and biomarker discovery. Statistical applications in genetics and molecular biology, 12(2), 207-223

Lee, S., Kim, J., Kang, H., Kang, D. Y., & Park, J. (2021). Genetic algorithm based deep learning neural network structure and hyperparameter optimization. *Applied Sciences*, *11*(2), 744.

Legler, D. F., Uetz-von Allmen, E., and Hauser, M. A. (2014). CCR7: Roles in cancer cell dissemination, migration and metastasis formation. *The International Journal of Biochemistry and Cell Biology*, 54, 78–82.

Li, B., Ren, B., Ma, G., Cai, F., Wang, P., Zeng, Y., ... & Deng, J. (2023). Inactivation of ZSCAN18 by promoter hypermethylation drives the proliferation via attenuating TP53INP2-mediated autophagy in gastric cancer cells. Clinical epigenetics, 15(1), 10.)

Li, C., Zhao, J., Kang, B., Li, S., Tang, J., Dong, D., & Chen, Y. (2024). Identification and validation of STAT4 as a prognostic biomarker in acute myeloid leukemia. Bioscience Reports, 44(2).

Liu, G., Li, J., Zhang, C. Y., Huang, D. Y., & Xu, J. W. (2021). ARHGAP20 expression inhibited HCC progression by regulating the PI3K-akt signaling pathway. Journal of Hepatocellular Carcinoma, 271-284

Liu, Y., Wang, X., Deng, L., Ping, L., Shi, Y., Zheng, W., Lin, N., Wang, X., Tu, M., Xie, Y., Liu, W., Ying, Z., Zhang, C., Pan, Z., Wang, X., Ding, N., Song, Y., and Zhu, J. (2019). ITK inhibition induced in vitro and in vivo anti-tumor activity through downregulating TCR signaling pathway in malignant T cell lymphoma. *Cancer Cell International*, 19(1), 32. https://doi.org/10.1186/s12935-019-0754-9

Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv Preprint arXiv:1705.07874*.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2019). *Consistent Individualized Feature Attribution for Tree Ensembles*. arXiv Preprint arXiv. http://arxiv.org/abs/1802.03888

Miao, M., Li, S., Yu, Y., and Li, F. (2023). LysM-containing proteins function in the resistance of Litopenaeus vannamei against Vibrio parahaemolyticus infection. *Developmental and Comparative Immunology*, asa

Nakamura, R., Oyama, T., Tajiri, R., Mizokami, A., Namiki, M., Nakamoto, M., and Ooi, A. (2015). Expression and regulatory effects on cancer cell behavior of NELL1 and NELL2 in human renal cell carcinoma. *Cancer Science*, 106(5), 656–664. https://doi.org/10.1111/cas.12649

Patel, S., Patel, H., Vyas, D., and Degadwala, S. (2021). Multi-Classifier Analysis of Leukemia Gene Expression From Curated Microarray Database (CuMiDa). *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, (pp. 1174–1178). Trichy, India

Quinlan, J. R. (2014). *C4. 5: Programs for machine learning*. Elsevier.

Rajasingh, J., Raikwar, H. P., Muthian, G., Johnson, C., and Bright, J. J. (2006). Curcumin induces growth-arrest and apoptosis in association with the inhibition of constitutively active JAK–STAT pathway in T cell leukemia. *Biochemical and Biophysical Research Communications*, 340(2), 359–368.

Roy, A., and Chakraborty, S. (2023). Support vector machine in structural reliability analysis: A review. *Reliability Engineering and System Safety*, 233, 109126.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

Rupapara, V., Rustam, F., Aljedaani, W., Shahzad, H. F., Lee, E., & Ashraf, I. (2022). Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model. Scientific reports, 12(1), 1000. https://doi.org/10.1038/s41598-022-04835-6

Sakamoto, Y., Ishida, T., Masaki, A., Murase, T., Ohtsuka, E., Takeshita, M., ... & Inagaki, H. (2022). CCR7 alterations associated with inferior outcome of adult T-cell leukemia/lymphoma under mogamulizumab treatment. Hematological Oncology, 40(5), 876-884

Saralamma, V. V., Vetrivel, P., Lee, H., Kim, S., Ha, S., Murugesan, R., Kim, E., Heo, J., and Kim, G. (2020). Comparative proteomic analysis uncovers potential biomarkers involved in the anticancer effect of Scutellarein in human gastric cancer cells. *Oncology Reports*, 44(3), 939–958. https://doi.org/10.3892/or.2020.7677.

Stagno, F., Russo, S., Murdaca, G., Mirabile, G., Alvaro, M. E., Nasso, M. E., ... & Allegra, A. (2025). Utilization of Machine Learning in the Prediction, Diagnosis, Prognosis, and Management of Chronic Myeloid Leukemia. International Journal of Molecular Sciences, 26(6), 2535.

Stricker, S. H., Köferle, A., & Beck, S. (2017). From profiles to function in epigenomics. Nature Reviews Genetics, 18(1), 51-66.

Sundaramurthi, H., Tonelotto, V., Wynne, K., O'Connell, F., O'Reilly, E., Costa-Garcia, M., Kovácsházi, C., Kittel, A., Marcone, S., and Blanco, A. (2023). Ergolide mediates anti-cancer effects on metastatic uveal melanoma cells and modulates their cellular and extracellular vesicle proteomes. *Open Research Europe*, 3.

Takasaki, Y., Iwanaga, M., Imaizumi, Y., Tawara, M., Joh, T., Kohno, T., Yamada, Y., Kamihira, S., Ikeda, S., and Miyazaki, Y. (2010). Long-term study of indolent ATLleukemia-lymphoma. *Blood, The Journal of the American Society of Hematology*, 115(22), 4337–4343.

Uchiyama, T., Yodoi, J., Sagawa, K., Takatsuki, K., and Uchino, H. (1977). ATLleukemia: Clinical and hematologic features of 16 cases. *Blood*, 50(3), 481–492.

Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

Wang, Y., Luo, Y., Fu, S., He, L., Pan, G., Fan, D., Wen, Q., and Fan, Y. (2023). Zinc finger and SCAN domain-containing protein 18 is a potential DNA methylation-modified tumor suppressor and biomarker in breast cancer. *Frontiers in Endocrinology*, 14, 1095604.

Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. Journal of Big Data, 11(1), 44.

Watanabe, M., Nakahata, S., Hamasaki, M., Saito, Y., Kawano, Y., Hidaka, T., ... & Morishita, K. (2010). Downregulation of CDKN1A in adult T-cell leukemia/lymphoma despite overexpression of CDKN1A in human T-lymphotropic virus 1-infected cell lines. Journal of virology, 84(14), 6966-6977.

Xie, S., Wu, H., Wang, Q., Cogswell, J. P., Husain, I., Conn, C., Stambrook, P., Jhanwar-Uniyal, M., and Dai, W. (2001). Plk3 functionally links DNA damage to cell cycle arrest and apoptosis at least in part via the p53 pathway. *Journal of Biological Chemistry*, 276(46), 43305–43312.

Xu, H., Jia, J., Jeong, H.-H., and Zhao, Z. (2023). Deep learning for detecting and elucidating human T-cell leukemia virus type 1 integration in the human genome. *Patterns*, 4(2).

Yamagishi, M., Nakano, K., Miyake, A., Yamochi, T., Kagami, Y., Tsutsumi, A., Matsuda, Y., Sato-Otsubo, A., Muto, S., and Utsunomiya, A. (2012). Polycomb-mediated loss of miR-31 activates NIK-dependent NF-κB pathway in adult T cell leukemia and other cancers. *Cancer Cell*, 21(1), 121–135.

Zhang, L., Zhou, L., Wang, Y., Li, C., Liao, P., Zhong, L., ... & Weng, J. (2022). Deep learning-based transcriptome model predicts survival of T-cell acute lymphoblastic leukemia. Frontiers in Oncology, 12, 1057153

Zhang, X., Wei, C., Liang, H., & Han, L. (2021). Polo-like kinase 4's critical role in cancer development and strategies for Plk4-targeted therapy. Frontiers in Oncology, 11, 587554.

Zhao, X., Lu, M., Liu, Z., Zhang, M., Yuan, H., Dan, Z., ... & Dang, C. (2023). Comprehensive analysis of alfa defensin expression and prognosis in human colorectal cancer. Frontiers in oncology, 12, 974654.).

Zhu, X. L., Zeng, Y. F., Guan, J., Li, Y. F., Deng, Y. J., Bian, X. W., ... & Liang, L. (2011). FMNL2 is a positive regulator of cell motility and metastasis in colorectal carcinoma. The Journal of pathology, 224(3), 377-388