Informatics in Medicine Unlocked 20 (2020) 100399

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

Treatment outcome classification of pediatric Acute Lymphoblastic Leukemia patients with clinical and medical data using machine learning: A case study at MAHAK hospital

Amirarash Kashef [a], Toktam Khatibi [a,*], Azim Mehrvar [b,c]

[a] School of Industrial and Systems Engineering, Tarbiat Modares University (TMU), Tehran, 14117-13114, Iran
[b] Mahak Hematology Oncology Research Center (Mahak-HORC), Mahak Hospital, Tehran, Iran
[c] AJA Cancer Epidemiology Research and Treatment Center (AJA-CERTC), AJA University of Medical Sciences, Tehran, Iran

ARTICLE INFO

Keywords:
Acute Lymphoblastic Leukemia (ALL)
children blood cancer
machine learning
treatment-related complications
MAHAK hospital
SVM
XGBoost

ABSTRACT

Introduction: Acute Lymphoblastic Leukemia (ALL) is the most common cancer among children. With the advancements of science and technology, the mortality rate of ALL is highly reduced. The aim of this study is treatment outcome classification of ALL patients aged less than 18 years with clinical and medical data using machine learning. For this purpose, ALL pediatric patients younger than 18 years treated at MAHAK multi-super specialty hospital from 2012 to 2018 are analyzed. Furthermore, MAHAK hospital is a reference center for treatment of childhood malignancies in Iran.

Data: In this study, data is collected manually from the paper-based records of 241 patients. Features included are patient demographic characteristics, medical information and treatment-related complications.

Method: Two scenarios are designed for data analytical purposes in this study. The first one considers all pediatric ALL patients but the second scenario excludes the patients with unknown cause of death from the study. As a whole, common classification algorithms are employed and tuned properly and compared to find the model showing superior performance.

Results: Our experimental results show that the XGBoost algorithm outperforms the compared classifiers with an accuracy of 88.5% (95% CI: 82.3–94.0) in the first designed scenario. On the other hand, the superior model in the second scenario is SVM with an accuracy of 94.90% (95% CI: 88.49–98.32) accuracy.

Conclusion: Despite several previous works that have analyzed gene expression data for ALL patients, the experimental results in this study show that clinical and medical data has reasonable importance in this area of research, too. Results show a significant improvement in the treatment outcome prediction utilizing the SVM algorithm. Moreover, our findings illustrate that the frequency of fever for a patient is the most predictive factor of the ALL treatment outcome.

1. Introduction

Blood is an important human body component that performs numerous vital functions, such as passing minerals, oxygen and carbon dioxide to the whole body to maintain metabolism. Blood has four essential components: Red Blood Cells (RBC), White Blood Cells (WBC), platelets (PLT) and Hemoglobin (HG) [1]. Leukemia is a type of blood or bone marrow cancer characterized by an irregular dramatic increase in the number of immature white blood cells named “blasts”. The term “leukemia” covers a wide spectrum of blood diseases [2]. Leukemia is classified into acute leukemia with a rapid progressive ability, and chronic leukemia that progresses slowly and has several obscure complications [1,3–5]. Acute leukemia infects the blood and bone marrow. Children and adults can develop numerous abnormal white blood cells in their body. Still, very recent developments have occurred to discover accurate preventive methods for acute leukemia disease [3]. Several risk factors have been identified for this dangerous and life-threatening disease. For example, the environmental factors such as exposure to benzene and ionizing radiation are highly associated with the development of childhood acute leukemia. Maternal history of fetal loss can also contribute to raise the risk of this fatal disease [6].

Acute type leukemia is classified into two classes based on a

* Corresponding author. Tel.: +982182883913.
E-mail addresses: amirarash.kashef@modares.ac.ir (A. Kashef), toktam.khatibi@modares.ac.ir (T. Khatibi), DRAZIMMEHRVAR@yahoo.com (A. Mehrvar).

https://doi.org/10.1016/j.imu.2020.100399
Received 16 April 2020; Received in revised form 14 July 2020; Accepted 15 July 2020
Available online 19 July 2020

French–American–British model, which is the most well-known classification model of leukemia including Acute Myeloid leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) [3,7,8].

Childhood ALL is a type of cancer that usually worsens quickly if it is not diagnosed properly and treated based on intensive chemotherapy protocols [9]. To provide the optimal care for pediatric patients, the oncologists not only must be familiar with the pathophysiology and treatment of childhood leukemia, but also must be aware of those latent complications that may occur during and after therapy, which could be fatal [10].

Cancer is the second most deadly disease in children, and leukemia is the leading cause of death in children [11]. Acute leukemia is one of the most common types of malignancy affecting the pediatric population. This malignancy constitutes about 30% of cancers occurring below the age of 15 years and 15% of cancers between the ages of 15 and 19 years. The pick incidence rate of pediatric acute leukemia fall within a 4-year age range, from 2 through 5 [12]. Indeed, boys are at a higher risk than girls in being diagnosed with ALL [6,12,13]. Therefore, this article focuses on ALL pediatric patients.

Treatment outcome is significant because it is the eventual result of all treatment modalities that have been used to treat a patient and predict whether a patient has survived from the disease or not. Indeed, it is vital for oncologists to find out the most important variables and factors to predict the treatment outcome. Accordingly, many investigations had been performed aiming to predict the long term outcome or treatment outcome utilizing gene expression data, and almost all of them had used data mining (DM) classifier algorithms, which shows the strength and importance of using machine learning methods for such purposes [14–16].

The previous studies have shown that ALL is largely documented amongst children. Age group is a contributing factor to the survival rate; overall, 85% of children and 50% of adults survive [3]. Vital promotion has been made in the treatment of pediatric leukemia over the past 70 years with current long-term survival rates of around 90% for ALL, compared to virtually 0% survival in the 1950s [17]. These statistics differ significantly for developing countries. For instance, in Iran, according to a meta-analysis [18], the 5-year survival in ALL patients for all ages is 57% (95% CI: 54.0–60.0) and for ages of 15 and below it is 61% (95% CI: 58.0–64.0). The age group that is selected for this research is [0, 17] which covers children from every age. We wonder if there are differences even between different age intervals amongst children and adolescents.

To the best of our knowledge, the most related works have used gene expression data for ALL treatment outcome prediction and classification. But, they have achieved low accuracies. In this study, a unique clinical dataset is collected and analyzed to predict treatment outcome for ALL pediatric patients. Additionally, the data used in this study contains information about treatment-related complications, and the novelty of this study is in considering the complication status as a class variable. The response variable has not only two labels, dead and survived, but instead, it has four labels which are dead because of a specific complication, survived but experienced different complications, survived and did not experience any complication, and dead because of an unknown cause, which has been named the unknown class.

The main aim of this study is to predict the treatment outcome of pediatric ALL patients based on classifying the clinical and medical data. ALL patients aged from three months old to 17 years at the time of diagnosis are classified in this study into four different classes by utilizing machine learning methods. For this purpose, two scenarios are developed; the first one considers the response variable having four classes but the fourth class (the unknown class) is excluded from the response variable in the second scenario. In both scenarios, data is analyzed with different machine learning algorithms varying from Decision Tree (DT), the most basic classification algorithm, to some complex classification algorithms such as SVM, Linear Discriminant Analysis (LDA), Multinomial Linear Regression (MLR), Gradient Boosting Machine (GBM), Random Forest (RF) and XGBoost.

The main novelties of this study lies in multiple folds including:

- The collected dataset in this study has a new combination of features for ALL treatment outcome prediction and our experimental results show it has reasonable accuracy in treatment outcome classification.
- Treatment outcome is basically formed into two labels (dead or survived) but we consider the treatment-related complications in four different classes.
- Several classification algorithms are compared to find the superior model for solving this problem and achieving the best performance.

The remainder of the paper is structured as follows, in section two related works are discussed, in section three is materials and methods, in which the data for the research is explained and preprocessing and the classification algorithms are described. Next, in section four the experimental results of this study are shown and finally, the discussion and conclusion is in the last section of the article.

2. Related works

Most of the previous related studies have used molecular or image datasets. Microscopic blood smear images have been analyzed for classifying different types of acute leukemia (ALL and AML). Furthermore, image datasets have been utilized to diagnose the mentioned diseases. On the other hand, gene expression data extracted from a DNA microarray has been widely used for different purposes namely, ALL detection [4,7], immunophenotype prediction [14], outcome prediction [15,16,19], ALL subtype classification [16,20,21], and relapse prediction [14]. Table 1 illustrates the summary of related prior works.

A previous study has proposed a hybrid method for microscopic image processing using powerful data augmentation for detecting B-cell lymphoblastic leukemia [20]. Moreover, another study has reviewed the previously proposed methods for detecting leukemia by image processing [24]. Gene-based classifications or gene expression data has been another interesting research topic in recent years. It is axiomatic that detection of leukemia subtypes mostly requires gene labels as it has been done with desirable accuracy of 97% in a previous study [21] utilizing 2D-clustering algorithms as well as supervised learning algorithms such as Decision Tree (DT), K-Nearest Neighbor (K-NN), Support Vector Machines (SVM) and Artificial Neural Networks (ANN). Articles reviewed in Ref. [3] mostly have used microarray datasets and blood smear images. Moreover, a previous study [25] has reported the intelligent techniques of molecular data analysis for leukemia in which the most reviewed studies have used microarray data, gene-expression data and/or the images. But, previous studies have shown that no specific genes are closely related to chronic leukemia [25].

The experimental results of the previous studies have shown that SVM often had the best performance. Moreover, some methods have been used effectively, for example for image based analysis of bone marrow samples. The Sequential Minimal Optimization (SMO) algorithm has been exploited to train a SVM classifier for feature selection and classification based on correlation for ALL diagnosis [22]. Feature extraction and classification of blood cells for an automated Differential Blood Count (DBC) system have been done utilizing the Multilayer perceptron (MLP), linear vector quantization (LVQ), K-NN, and SVM [23].

Furthermore, some investigations have been done to predict the immunophenotype, outcome and relapse based on gene expression data. For instance, a previous study [14] has predicted the outcome and relapse after treatment by gene expression data with the accuracy of 74% and 87%, respectively.

Another previous study has considered 99 children with high-risk ALL to find the predictive genes of early response and long-term outcome. For this purpose, linear regression with the LOOCV approach has been used with an accuracy of 75% and area under the

Table 1
Summary of the related prior works.

Reference	Problem	Dataset	Preprocessing	Analytical method	Remarks
[1]	Classifying blasts of acute leukemia into ALL and AML	Blood smear images	–	KNN (K = 4)	They have reached the accuracy up to 80%
[2]	Distinction between ALL and AML	Gene expression	SMIG (selecting the most informative genes)	Enhanced Classification algorithm (ECA)	The algorithm has reached the accuracy of 98%
[4]	ALL detection	Blood smear images	Two-stage color segmentation based on a fuzzy clustering method	SVM classifier	The final performance has been validated based on a hematologist opinion
[8]	Blasts classification into ALL and AML	stained peripheral blood smears	segmentation, feature extraction	Several classification algorithms	Best performance has obtained with SVM (accuracy of 92%)
[9]	Management of ALL	Clinical data	Missing value replacement, noisy data and inconsistent data detection	Decision trees	–
[14]	Prediction of immunophenotype, treatment response, and relapse in pediatric ALL	Gene expression	Outlier detection, feature selection with PCA	KNN, ML, NC, NSC, LDA, and SVM	Low accuracy for long term outcome and relapse prediction have been obtained
[15]	outcome prediction based on the Genomics of Leukemia	Four different datasets, cDNA, Affymetrix, SNP and clinical data	Missing value replacement with a constant value of zero	RF, SVD, SVM	The obtained results have shown no robustness for the classifiers
[16]	Classification, subtype discovery, and prediction of outcome in pediatric ALL	Gene expression	Gene extraction with CFS approach from DNA microarray	H-clustering, SVM, PCA, ANN, SOM, all with LOOCV	Perfect accuracies have been obtained for subtype classification and relapse prediction
[19]	Identifying pediatric ALL at initial diagnosis who are at risk for inferior response to therapy	Gene expression	Gene extraction using AffymetrixMAS 5.0 Software.	Logistic regression with LOOCV	Significant predictive genes for early response and long term outcome have been identified
[20]	Distinguishing between immature leukemic blasts and normal cells only for B-lymphoblastic leukemia	blood smear microscopic images	Normalization, resizing and data augmentation	Deep learning-based method specially convolutional neural network	Robust method has been proposed with the accuracy of 96.17%; however, AUC has not been reported
[21]	identifying the known prognostic subtypes of ALL	Gene expression	Data separation and feature extraction	PCA, KNN, SVM, ANN	They have reached the accuracy of about 97%
[22]	classifying a lymphocyte as normal or blast	blood microscopic images of lymphocyte	Segmentation, feature extraction and feature selection	Sequential Minimal Optimization algorithm to train an SVM classifier	They have extracted a set of 16 features and achieved the accuracy of 92.3%
[23]	Automatic blood cell detection and counting	Blood cell images	Segmentation and feature extraction	KNN, LVQ, MLP, SVM	They have obtained the accuracy of 91% using SVM, the second best performance has been reached with MLP with the accuracy of 89.7%

Relative Operating Characteristic (ROC) curve of 80% which have been obtained for long-term outcome prediction [19].

To the best of our knowledge and by considering the previous studies, such as review articles [3,25], we find out that the main focus of the previous studies has been proposing and using DM methods for automatic detection and diagnosis of leukemia subtypes, and very little attention has been done for classifying treatment outcome. However, treatment modalities prescribed and followed for a specific patient and the complications occurring during the therapeutic procedures can influence outcome. Therefore, this study uses data mining and machine learning methods to predict the treatment outcomes for children suffering from ALL. For this purpose, clinical and medical data is analyzed in this study.

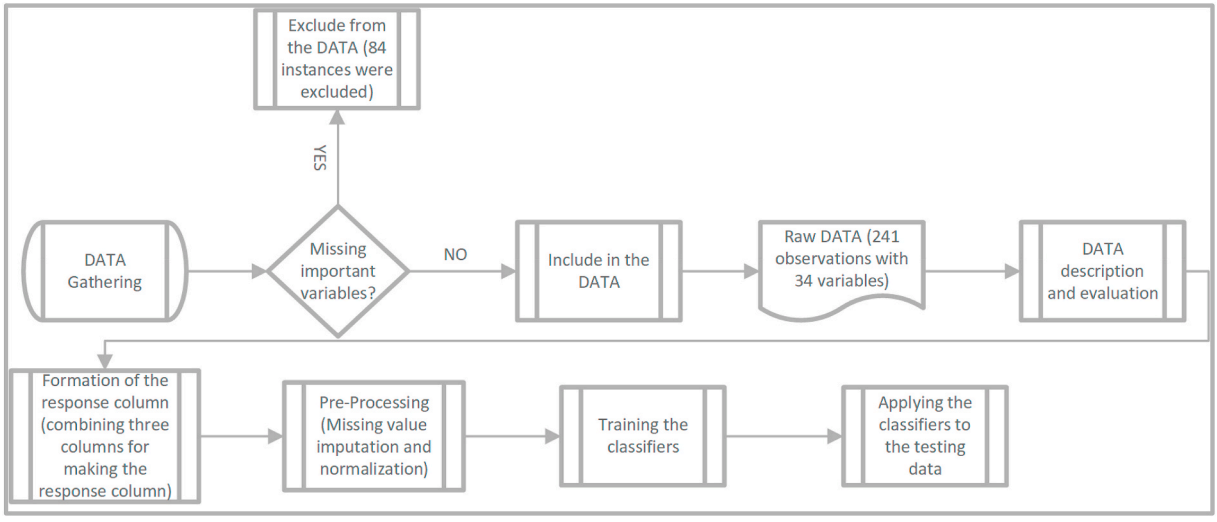


Fig. 1. The main steps of the research methodology in this study for ALL treatment outcome prediction.

3. Materials and methods

The main steps of the research methodology in this study are shown in Fig. 1 based on CRISP-DM methodology [26], which with more detail will be described in the following subsections.

In this study, we are aiming to find the best classification performance on a clinical and medical dataset. This dataset is collected by ourselves from the medical paper-based records of the ALL patients who have used the leukemia therapeutic procedures at the MAHAK pediatric cancer treatment hospital.

3.1. Data description and evaluation

The retrospective clinical and medical data for this study is gathered from the paper-based records of 241 ALL patients from 2012 to 2018 from MAHAK's Pediatric Cancer Treatment and Research Center (MPCTRC). This dataset includes 144 Male (60%) and 97 Female (40%). MAHAK's Pediatric Cancer Treatment and Research Center (MPCTRC) is one of the main national referral centers for childhood malignancies. Due to the large number of referrals, data compiled at this center can be considered as reference for any issues related to national health strategies and policies in order to facilitate and optimize the medical services for pediatric malignancies [27].

For each patient admitted to MAHAK hospital, some information including blood tests, clinical trials and transcriptions are recorded and kept confidential in two files including clinical and inpatient files. Each time the patient is hospitalized, a new record describing the current clinical and medical features for the patient is added to the corresponding inpatient file.

Clinical and inpatient files for every 241 patients are thoroughly read and analyzed. In this way, the values of 31 variables are collected including four major blood components (WBC, RBC, HG, PLT) of the first blood test with which the group of pediatricians and oncologists in MAHAK hospital diagnosed ALL for the patient, risk group, cell lineage either it is B-cell or T-cell, treatment modalities and 17 different complications during therapy or treatment-related complications which will be discussed in the next section. Table 2 illustrates the acquaintance with all attributes included in the data regarding their types and values.

The mean ± STD of age for patients is 6.56 ± 4.38, ranges from 3 months to 17 years. Fig. 2 shows the distribution of the patients' age at the time of diagnosis.

As illustrated by Fig. 2, the most incidences of ALL occur at the age of 2 and 3 years old, with 34 patients (14.1%) which is mentioned in the previous studies [6,12,27–29], too. ALL manifests in the big proportion of children between the age of 2–5 years and is more frequent in males than females. Furthermore, all of the patients witnessed intensive chemotherapy, 27 patients (11.2%) have been prescribed radiotherapy. 14 patients (5.8%) have experienced a situation that a group of pediatricians and oncologists in MAHAK hospital unanimously have decided that Allogeneic Hematopoietic Stem Cell Transplant (Allo-HCST) have been required. But, unfortunately, six of those have died after transplant.

Appendix 1 shows the initial WBC count scaled on the age at the diagnosis time which the maximum value is 284,000 for a 12-year old boy and the lowest value is 500 for a 15-year old boy.

Appendix 2 illustrates the initial RBC count again scaled on the age at the diagnosis time reaching top at 5,640,000 counts for a three-year old boy and hit a low at 274,000 for a two-year old boy.

Appendix 3 shows the initial PLT count scaled on the age at the diagnosis time, top value is 955,000 for a four-year old boy and the lowest count is 5000 for a two-year old boy; surprisingly, for the same boy that has got the lowest RBC count.

Appendix 4 compares the initial HG count scaled on the age at the diagnosis time which the maximum count is 16.2 $\frac{gr}{dlitr}$ for a 15-year old girl and the minimum count is 2.4 $\frac{gr}{dlitr}$ for a two-year old girl.

Table 2
Describing the types and values of the variables considered in our collected dataset.

Variable	Type	Value
Gender	Binary (Boy and girl)	"1" = Boy, "0" = Girl
Age at diagnosis time	Numeric	Min = 0.25, Max = 17
WBC (White Blood Cells)	Numeric	Min = 500, Max = 284,000
RBC (Red Blood Cells)	Numeric	Min = 274,000, Max = 5,640,000
PLT (Platelets)	Numeric	Min = 5000, Max = 955,000
Cell type	Binary (B-cell and T-cell)	"1" = B-cell, "0" = T-cell
HG (Hemoglobin)	Numeric	Min = 2.4gr/dlitr, Max = 16.2gr/dlitr
Risk Group	Ordinal (Mild, Moderate, high)	"1" = Mild, "2" = Moderate, "3" = high
Radiotherapy	Binary	"1" = Yes, "0" = No
Allogeneic Stem Cell Transplant (Allo-SCT)	Binary	"1" = Yes, "0" = No
L-Asparaginase	Binary	"1" = Yes, "0" = No
Refractory	Binary	"1" = Yes, "0" = No
Relapse	Numeric	Min = 0, Max = 3
Thrombosis	Binary	"1" = Yes, "0" = No
Pulmonary infection	Binary	"1" = Yes, "0" = No
Pulmonary failure	Binary	"1" = Yes, "0" = No
Fungal infection	Numeric	Min = 0, Max = 2
Kidney failure	Binary	"1" = Yes, "0" = No
Transient hyperglycemia	Binary	"1" = Yes, "0" = No
Mediastinal mass	Binary	"1" = Yes, "0" = No
Pancytopenia	Numeric	Min = 0, Max = 5
Convulsion	Numeric	Min = 0, Max = 2
Herpes	Numeric	Min = 0, Max = 2
Pancreatic cyst	Binary	"1" = Yes, "0" = No
Gastroenteritis	Binary	"1" = Yes, "0" = No
Fever	Numeric	Min = 0, Max = 12
Immunocompromised condition	Numeric	Min = 0, Max = 50
Pneumonia	Numeric	Min = 0, Max = 2
Neutropenia	Numeric	Min = 0, Max = 6
GVHD	Binary	"1" = Yes, "0" = No

Common type of ALL is B-cell and based on our data this fact is obvious as there are 216 patients (89.6%) and 25 patients (10.4%) suffering from the B-cell and T-cell type, respectively. This variable is extracted from immunophenotyping test.

Fig. 3 demonstrates the proportion for each risk group stratification, according to patient's clinical file. The patients are classified into three risk groups including Mild Risk (MiR), Moderate Risk (MoR) and High Risk (HR). 142 patients (59%) lie in MiR class, 78 patients (32%) are classified into MoR and 21 patients (9%) lie in the high risk requiring intensive care. The value of this variable is determined by a group of oncologists.

Outcomes in our data show a 10% death after the treatment procedures and 219 (90%) patients survive the fight with their disease. It shows the powerful and accurate treatments are prescribed and followed in MAHAK hospital. Besides, the disease relapse is an important feature which would complicate the treatment procedure for a patient suffering from ALL. In our dataset, the relapse frequency recorded for three patients has been three times and eventually they all have died.

Lastly, Appendix 5 shows the pairwise scatterplots of the data which are colored based on the four levels of the response variable by RStudio software in order to obtain a better understanding and perception of the relationships between the numeric variables.

The major proportion of the dataset analyzed in this study includes the important and common complications that ALL pediatric patients confront during cancer therapy. Therefore, these features will be explained thoroughly and some statistics are presented in the following subsection.

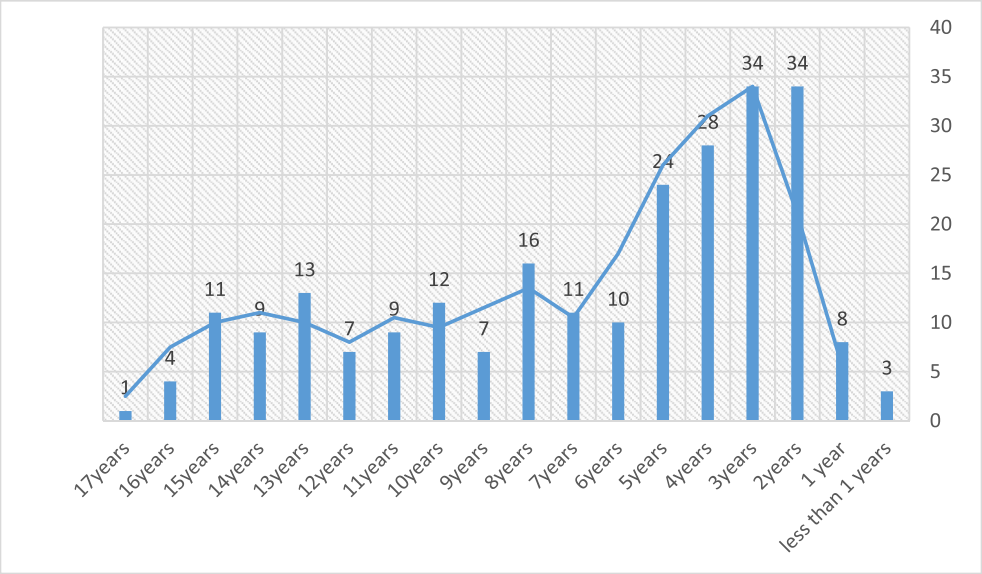


Fig. 2. Patient's age distribution at the time of diagnosis.

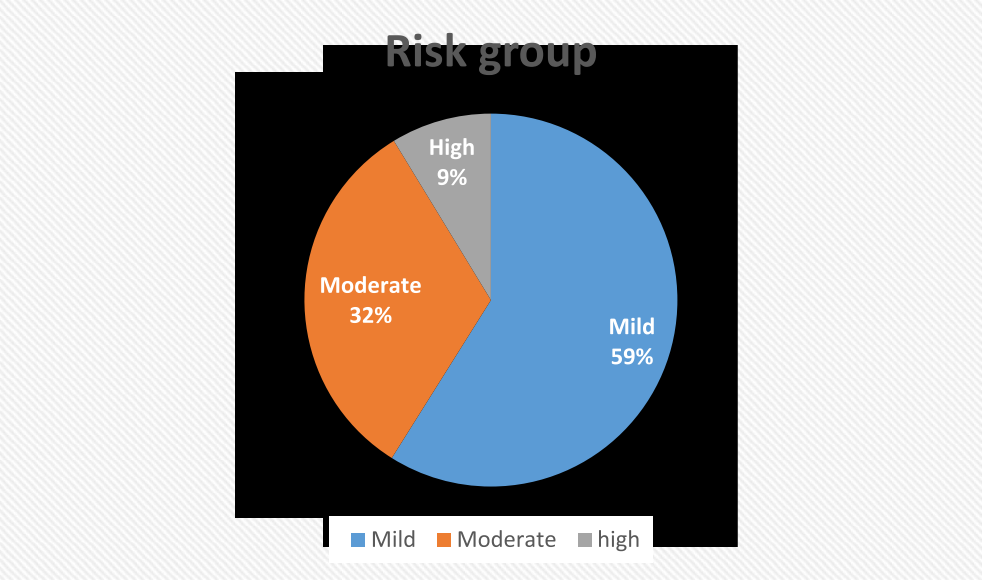


Fig. 3. Proportion for each risk group according to the data.

3.1.1. Treatment-related complications for ALL patients

Physicians and pediatricians should be aware of the possible complications that may manifest during the therapy due to both the disease and its treatment. Since the patients suffering from ALL are immunocompromised due to both the disease and its treatment, they are at high risk for infection. The pediatrician caring for these children must be aware of common infectious complications and their management [10]. In most patients, however, chemotherapy itself is the main cause of occurring neutropenia. Many patients undergoing anti-leukemic treatment will have multiple episodes of chemotherapy-induced neutropenia [10].

The hallmark of infection in a neutropenic child is fever, either a single temperature of 38 °C or greater. The presence of fever in a neutropenic child requires immediate attention because infection may be rapidly progressive; delaying treatment until culture results are obtained may be fatal [10]. Most of the epidemiological studies have revealed the symptomatic features of the patients enrolling with ALL like bone pain, fever, organomegaly, nausea and anorexia [27]. Although, there are vital features and symptoms for a child in diagnosing the disease, in this research our focus is mainly on the complications occurring during the cancer therapy which might be fatal if they are not treated well.

Pneumonia is a pulmonary infection and is highly common in children with leukemia. The previous studies have revealed that pneumonia constitutes for 28 to 43% of the fatal and nonfatal infections amongst these patients [30].

Venus Thrombosis (VT) can be a relatively common complication of treatment in children with acute leukemia. A meta-analysis have demonstrated that the symptomatic thrombosis has occurred for 5.2% of the studied population [17,31].

Anterior mediastinal masses are characteristic of T-cell ALL and have anticipated to occur in 53–64% of pediatric patients [17,32]. Accordingly, in our case there are three patients (12%) out of 25 patients with T-cell lineage that caught mediastinal mass and all of them survived due to high-quality healthcare services in MAHAK hospital.

The risk of bleeding complications in pediatric patients with acute leukemia on anticoagulation is significantly low at 2%, and therefore often the benefit of preventing further thrombosis may outperform the potential risk of bleeding [31].

Between 2 and 18% of patients who are treated with Asparaginase develop acute pancreatitis. Pancreatitis is a cause of considerable morbidity. Asparaginase as an essential chemotherapeutic agent in the treatment of ALL is available in three formulations (native *E. coli* L-asparaginase, peg-asparaginase, and *Erwinia* L-asparaginase). It works with reducing plasma concentrations of asparagine by metabolizing it into aspartic acid and ammonia.

This dearth of asparagine results in the deprivation of this amino acid in the leukemic blasts resulting in cell death. Therefore, the physicians increase the dosage of L-asparaginase in order to kill as many of the leukemic cells as possible. But, in some cases, the patients show an allergy to the drug. Thus, the physicians change the L-asparaginase to Peg-asparaginase. Our collected dataset includes this hidden variable indicating that 12 patients (4.97%) have suffered from an allergic reaction to the L-asparaginase [17,33].

The graft-versus-leukemia effect is one of the most important biological effects influencing the survivability in patients with acute leukemia. The recognition of this modality over the past three decades has led to far-reaching changes in the concept and conduct of allogeneic stem cell transplant in ALL patients, and in the infusion of donor lymphocytes as a therapeutic modality [34].

We have identified and extracted 17 common complications from the inpatient files including thrombocytopenia, pulmonary infection, fungal infection, transient hyperglycemia, mediastinal mass, pancytopenia, convulsion, herpes, pancreatic cyst, gastroenteritis, fever, pneumonia, neutropenia and immunocompromised condition. The immunocompromised condition is interpreted from the transcriptions that were written down, “Hospitalize the patient in the isolated room because of immunocompromised condition”. Besides, there were four complications which induced death, kidney failure, pulmonary failure and Graft-Versus-Host Disease (GVHD) and in one case thrombosis caused mortality.

All of these complications are diagnosed by a group of professional physicians in MAHAK hospital. The frequency of experiencing a specific complication for every patients is considered in this study, too.

3.2. Pre-processing

Heterogeneity of the paper-based medical records creates challenges for patient health data analysis. Since different patients have their own complications and disease status, clinicians and pediatricians would decide different treatment plans for each patient. This irregularity in the patient health data leads to a lack of structure that complicates the feature learning and classification tasks [35].

In our work, data preprocessing is done during data gathering by ignoring those patients that has got low rate of information trying to result in an integrated data. After data collection, missing value handling and data and normalization tasks are required.

But, before missing value imputation, the training and test datasets should be sampled and separated. Therefore, 10-fold cross validation (C. V.) is used to divide the dataset into training and test datasets. Then, training dataset is partitioned into with ratio of 60:40 to form training sample for training the classifiers and validation sample for tuning the hyperparameters of the models and avoiding the overfitting of the classifiers. For this purpose, some values for each hyper-parameter of the classifiers are pre-specified and Grid search method is used for tuning the hyperparameters. Each classifier with a pre-specified hyper-parameter values is trained on the training sample. Then, its performance is evaluated via applying it to the validation sample. For identifying which models are prone to overfitting, the performance of the classifiers is computed and compared for training sample and validation sample. If the performance of a model is more desirable for training sample than validation sample, it may show that the model is overfitted.

3.2.1. Missing value imputation

Real-world datasets have some missing values. Therefore, the previous studies have tried to find the best imputation strategy [36]; however, since we ourselves gathered the data, there are a few missing values in our collected dataset. The most missing value rate is related to RBC count that is 0.145. Only four variables have got missing values so they are simply imputed with missForest package in RStudio [37] which leads to 0.132 out-of-bag error OOB (error).

3.3. Response variable formation

In this study, the treatment outcome for ALL patients will be classified into four groups. Therefore, we encounter with a multi-class classification problem. We assume that all the patients have reached the end of their treatment. Table 3 shows the construction procedure of the response variable based on three different variables.

When the data were being collected, in the raw version of the data we have got three attributes named Complication (valued 0 or 1), Leukemia treatment success (dead (value = 0) or alive (value = 1)) and Complications treatment success (dead because of a treatment-related complication for example GVHD (value = 0) or alive after treating complications (value = 1)). Next, we formed our response column with combining these three variables according to Table 3. A patient is a member of class1 if he or she has died because of a complication which is one of the components of this list (Pulmonary failure, Kidney failure, Thrombosis and GVHD). A patient is a member of class2 if he or she is alive after the therapy which means that patient has survived without any serious complication during therapy. A patient is a member of class3 when he or she has survived from leukemia and had experienced some complication, and finally, a patient is a member of class4 when he or she died during therapy but we don’t precisely know what was the main cause of his/her death. The fourth-class patients did not die in the MAHAK hospital, thus a record was not available and we call it the unknown death class. As a whole, class1 and class4 indicate the patients that have died and class2 and class3 indicate the patients who have survived from the disease.

Due to the great survivability for children from leukemia, it is axiomatic that the majority of the classes corresponding to the dead population is very low. Therefore, as shown by Fig. 4, this study encounters with an imbalanced classification task. The majority classes are class2 and class3. On the other hand, class1 and class4 are our minority classes because the dead population after therapy are very low. Classifying imbalanced datasets can lead to overfitting the models. In order to avoid this inconsistency, all evaluation metrics are calculated and compared on both training dataset and validation dataset to track the overfitting phenomenon.

3.4. Classification algorithms

For classifying our collected data, several different classifiers are used in this study. First classification algorithm and the most descriptive one is the decision tree and it is the basic algorithm for many ensemble classifiers which some of those are included in our list of comparison.

Table 3
The formation of the response variable.

Complication (0or1)	Leukemia treatment success (0or1)	Complications treatment success (0or1)	Response column: patient’s outcome status
1	0	0	1
0	1	Empty	2
1	1	1	3
1	0	1	4

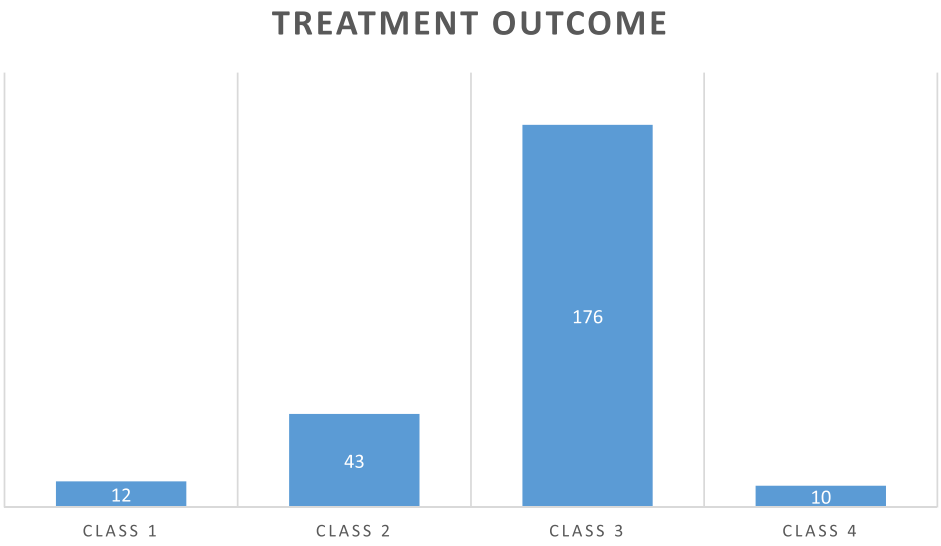


Fig. 4. Abundance of the four classes in the response column (sum of class1 and class4 equals just 9.12% of the whole response column and only class3 accounts for 73.02% of the response column).

SVM is one of the most common algorithms utilized in many data mining researches. Random forest (RF) is one of the ensemble classifiers based on decision tree and is a powerful tool for classification.

LDA [38] is the first method developed for multidimensional classification perspective and has been used for several years as the main classification technique.

Logistic Regression (LR) is a widely used technique for pattern recognition, sometimes being used for classification problems with discrete response variable; however, we cannot make proper use of LR, thus multinomial logistic regression which is a new approach for multiclass classification substituting the LDA approach.

Other models that are being used effectively in our research are Gradient Boosting Machine (GBM) and eXtreme Gradient Boosting (XGBoost). Multilayer Perceptron (MLP) as a neural network model is also implemented, but leads to a poor performance. Therefore, MLP is excluded from the models compared in this study.

DT is a popular classifier which can extract a tree-like diagram and can be understood and interpreted by all persons without requiring to be familiar with data mining concepts. It has not the black-box nature. Moreover, it has high speed to converge. Another advantage of DT is that pre-pruning and post-pruning techniques can be used to prevent from overfitting. But, its main drawback is that it shows desirable performance when data is linearly-separable.

SVM is a strong classifier which leads to highly desirable performance in many applications. Using linear kernels, SVM can classify only linearly-separable data. But, while using nonlinear kernels such as polynomial, radial basis and sigmoid kernels, SVM can classify nonlinearly-separable data with high performance, too. To avoid it from overfitting, the cost coefficient in the objective function for optimization should be tuned carefully.

RF is an ensemble of several DTs and it is a strong classifier, too. RF can classify both linearly-separable and non-linearly separable data with high accuracy. RF is one of the fastest ensemble classifiers, too. To prevent from overfitting, the number of DTs, the maximum depth of each DT, the minimum leaf size and some other hyperparameters can be tuned.

MLR has been widely used in the medical applications specially for simultaneously predicting multiple output variables. The advantages of classifying data using MLR are that this may be useful for achieving to a more accurate understanding of the relationship between each input variable with the output.

LDA uses information from all input variables to construct a new artificial variable to minimize the variance and maximize the class distance among the output values.

GBM have shown desirable performance in many different applications. Whereas RF is a bagging ensemble of independent DTs, GBM trains an ensemble of shallow and weak successive DTs with boosting method to improve the performance of the next DTs compared to the previous ones in the ensemble. Therefore, GBM show high accuracy while classifying data and often, other classifier cannot beat it.

XGBoost is one of the top machine learning classifiers and is widely used and popular classifier. It is a highly flexible and versatile tool for classification with the ability of user-built objective functions.

3.5. Evaluation metrics

As mentioned in the previous subsection, different classification algorithms are implemented using Rstudio in this study. To perform a fair comparison between the classifiers and finding the superior model with the best performance, several evaluation metrics which are well-suited for our case and our unique dataset are considered and calculated. Our considered evaluation metrics include accuracy (Eq. (1)), precision (Eq. (2)), recall (Eq. (3)), and the most important one which is F1-score (Eq. (4)) as a necessary measure for evaluating an imbalanced multiclass classification. Another significant evaluation metric which is used in our research is the Area Under the ROC Curve (AUC). Some of these measures are suitable for binary classification tasks. Therefore, multi-class version of these measures such as multi-class AUC are used in this study [39]. Since, we encounter to an imbalanced classification task, overfitting phenomenon is highly possible and should be strictly monitored and avoided.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = 2^{*}\frac{Precision^{*}Recall}{Precision + Recall} \tag{4}$$

4. Experimental results

Our experimental results are divided into two scenarios. The first scenario considers the response variable with four labels and the second one excluded the unknown death cause class (the fourth class) based on several reasons. Firstly, unknown death cause may bias the classification results. Secondly, this class has got the fewest number of patients. The last but not least reason is that the most misclassification in the first scenario is related to the unknown death cause class or fourth class and almost every model is overfitted. Several efforts for tuning hyperparameters do not improve the performance of the models in the first scenario significantly. According to our experimental results, the only classification algorithm that is not overfitted in the first scenario with superior performance, is XGBoost model.

For all classifiers, tuning the hyperparameters are performed with Grid search method in this study to reach the best performance for treatment outcome prediction for ALL patients. Grid search method is one of the methods which have been used in the previous studies for tuning the hyperparameters. A main advantage of this method is its high speed among the compared methods which have been proposed for this purpose in the previous studies.

Table 4 illustrates the evaluation metrics on test dataset for XGBoost model. Table 5 shows the confusion matrix for XGBoost model as the superior model leading to the best performance in the first scenario.

For the second scenario, the unknown death cause class (the fourth class) is omitted. It leads to remaining 231 instances which are partitioned with 10-fold CV strategy. The hyperparameters of the classifiers are tuned similar to the first scenario. Table 6 compares the performance of different classifiers in the second scenario.

As illustrated by Table 6, SVM (with cost of 100) is the superior model leading to the best performance with the accuracy of 94.90% (95% CI: 88.49–98.32) and the multi-class AUC of 87.7% and p-value of 3.206×10^{-7}. These statistics show the significant power of the considered variables in this study for classifying and predicting the treatment outcome for ALL patients. Table 7 is the confusion matrix corresponding to SVM classification results obtained on the test dataset.

As illustrated by Table 7, only five observations out of 99 cases are misclassified.

Finally, for more detailed and comprehensive results, ROC curves of the four best algorithms of the second scenario including SVM, RF, DT and MLR are shown in Fig. 5.

4.1. Feature importance and scoring

Experimental results show that RF and SVM are our two best performed classification algorithms. Thus, feature importance is extracted based on the two algorithms. Firstly, random forest gives the feature importance and Fig. 6 shows the feature ranking based on RF with two scales, accuracy and Gini Index.

As illustrated by Fig. 6, both accuracy and Gini index vote the fever as the most important variable and neutropenia as the second most important variable for the treatment outcome classification. Also, the coefficients derived from SVM indicates that the fever variable is identified as the most important variable for predicting treatment outcome for ALL patients, too.

Table 4
The best result, maintained from the XGBoost model for the first scenario.

Evaluation metrics	Accuracy	Precision	Recall	F1-measure	Multi-class AUC
XGBoost model	88.54%	62.31%	90.23%	73.72%	0.79

Table 5
The confusion matrix for the XGBoost model.

	Class 1 (actual)	Class 2 (actual)	Class 3 (actual)	Class 4 (actual)
Class 1 (predicted)	2	0	0	0
Class 2 (predicted)	2	11	3	0
Class 3 (predicted)	2	0	71	0
Class 4 (predicted)	0	0	0	1

Table 6
Comparing the performance of different classifiers in the second scenario.

	Accuracy	Precision	Recall	F1-measure	Multi-class AUC
DT	86.73%	61.76%	54.09%	57.67%	78.43%
SVM	**94.90%**	**90.23%**	86.22%	**88.17%**	87.7%
RF	90.91%	79.61%	**92.17%**	85.43%	79.8%
MLR	91.84%	80.05%	85.77%	82.81%	77.15%
LDA	85.86%	64.38%	71.49%	67.75%	70.06%
GBM (over-fitted)	83.84%	51.74%	54.29%	52.99%	66.38%
XGBoost (over-fitted)	86.87%	63.74%	74.82%	68.84%	71.60%

Table 7
The confusion matrix for the SVM model.

	Class 1 (actual)	Class 2 (actual)	Class 3 (actual)
Class 1 (predicted)	4	0	2
Class 2 (predicted)	0	18	1
Class 3 (predicted)	1	1	71

5. Discussion

The main focus of this study is classifying patients based on their treatment outcome. This research problem has been considered previously by analyzing gene expression, DNA microarray structure or medical images. To the best of our knowledge, it is the first time that clinical and medical variables collected from paper-based medical records are used for predicting the treatment outcome for ALL patients. Another advantage of this study is that the data collection does not require enhanced technologies like those required for extracting DNA microarray, so the variables are measured as the events happen. This would luckily shine a new path of research in this area.

Feature importance and ranking is very significant especially in our study where the data is intact. Some of the classifiers such as RF and SVM with linear kernel are used as the feature ranking methods simultaneously. Based on our experimental results of feature ranking with RF and SVM, it is concluded that the most important feature for treatment outcome prediction is the fever variable. This variable counts the frequency of catching fever during the time of the leukemia treatment for a patient. The domain experts at MAHAK hospital consider this finding as a valuable point.

It is the first time that this kind of variables are included in the dataset and resulted in a high accuracy. In particular, there is a variable called immunocompromised condition which is extracted from the prescription where the correspondent physician wrote “hospitalize the patient in the isolated room because of immunocompromised condition”.

Treatment-related complications which are gathered for this study include low risk and high risk complications. An example of a low risk complication is mediastinal mass for T-cell patients or herpes. A high risk complication includes GVHD that may induce death or kidney

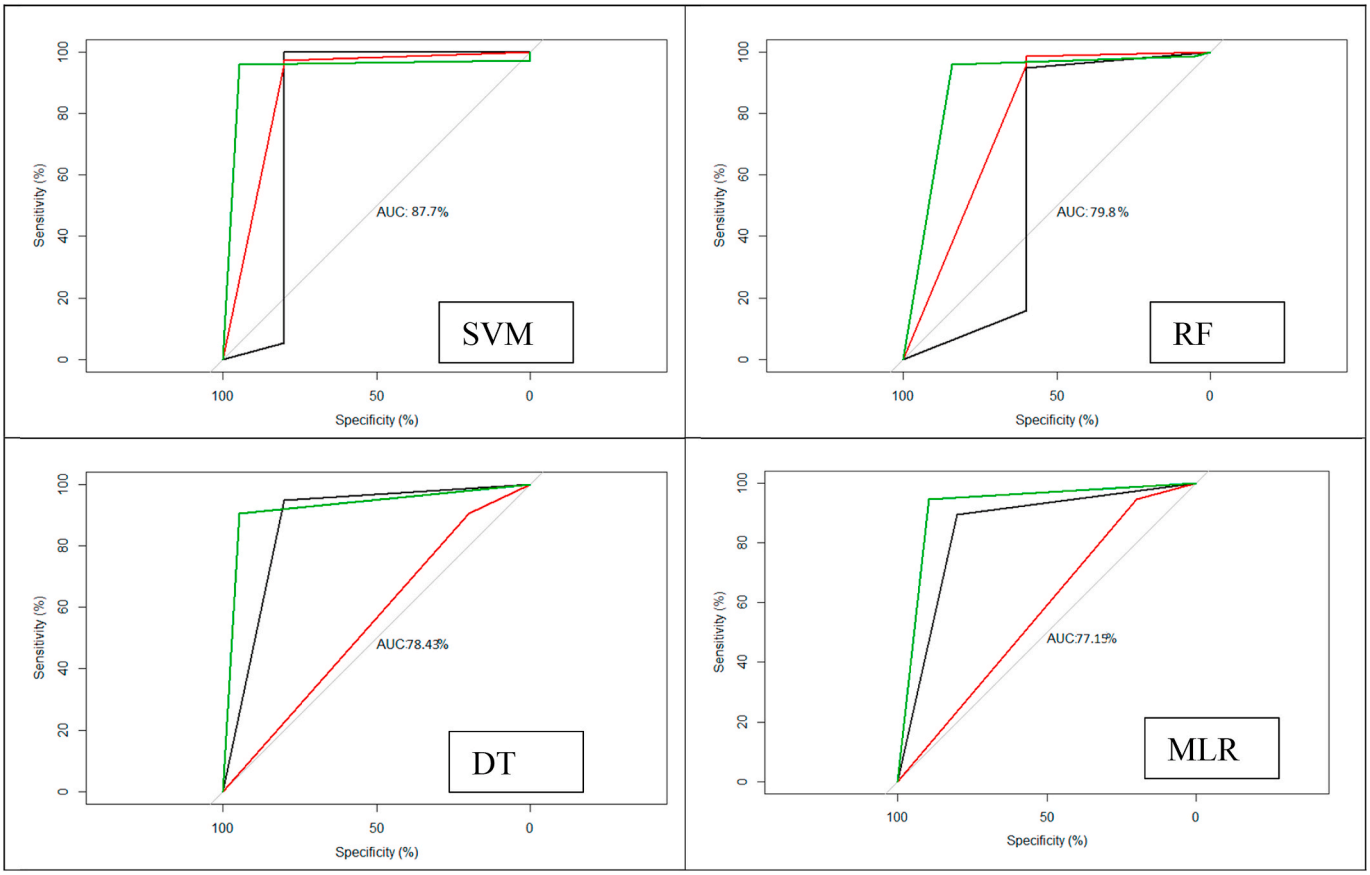


Fig. 5. ROC curves of the four best algorithms in the second scenario.

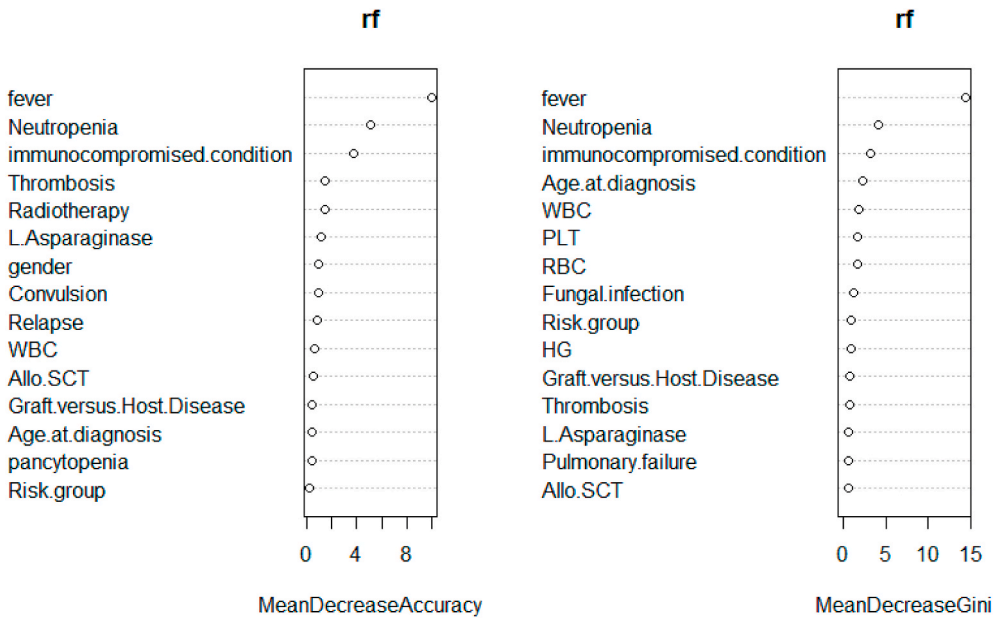


Fig. 6. Feature ranking extracted by RF.

failure and most definitely will cause death. Even though high risk complications may cause mortality, GVHD takes 3–6 months to show its extreme effects [34]. In the present study there are four complications that according to patients' files caused death; however, a considerable time interval exists from occurring the cause of a complication, showing its symptoms, its diagnosis and mortality due to the complication. Furthermore, almost every kind of treatment-related complications can be treated successfully if it is early diagnosed. Precise calculation and further analysis of the time interval between measuring clinical features and observing the actual outcome and early diagnosis of treatment-related complications can be an interesting future research opportunity.

In this technological era that gene expression solutions are the main insight into advancements, this old-fashion data gathered from paper-based records were somehow taken the lead for this particular issue. Moreover, expanding the labels for the treatment outcome status from just considering death or alive to more detailed labels has opened a new doorway for the researchers to reconsider some concepts for keeping more comprehensive and precise results.

Machine learning is the most widely used technique in classification and prediction problems. Most of the studies consider some of the major machine learning algorithms and only report the best performance and do not tend to draw a comparison between different algorithms, whereas, in this study we have drawn a comparison between seven most common classification algorithms and determined the superior model with the best performance. Similar to a wide proportion of related works, SVM leads to the best accuracy for treatment outcome prediction in this study.

Furthermore, in almost every study investigated until this date that included data, that class of patients which is referred to as unknown were excluded in the first place but we considered those in a separate scenario and compared the results with and without that group of patients.

Generally speaking, there were some limitations and difficulties regarding this actual study. Firstly, the major restriction was about the data gathering which we had to deal with paper-based records and many inconsistencies that has emerged during the data gathering. Another issue in this regard was the huge number of papers and files which should be read in order to retrieve the data. Last but not least, because this data is unique there were a lot of alterations in the features and every little change required an agreement of the group of experts.

Finally, despite low survival rate in Iran as a whole, MAHAK hospital is a pioneer in treatment of childhood diseases specially leukemia. That's all the reason why we have got few number of deaths in our data; however, with the knowledge of having few number of deaths in our data which represent our death class, we did get reasonable performance for our classifiers. It is completely acceptable that we should have gather more observations from dead population and we did actually but as we mentioned there were 84 observations that we were forced to omit due to the lack of important features including 27 dead patients which we could make our results robust with. An important evaluation metric used in this study is F1-measure which is rarely used in this kind of studies. But it is used because of imbalanced number of observations in each class of the response column.

6. Conclusion

In this study, we aimed at treatment outcome classification for 241 ALL children and adolescents aged between three months and seventeen at the time of diagnosis. Dataset for this study were gathered manually from paper-based records in the MAHAK hospital, which is specialized for childhood diseases, by ourselves. Two scenarios were considered for the methodology. In the first scenario, the tuned XGBoost showed the best performance with an accuracy of 88.54%, precision of 62.31%, recall of 90.23%, F1-measure of 73.72% and multi-class AUC of 79%. Meanwhile, SVM with the cost parameter of 100 outperformed the compared models for the second scenario with an accuracy of 94.90%, precision of 90.23%, recall of 86.22%, F1-measure of 88.17% and multi-class AUC of 87.7%, which is a desirable performance for a classification task.

Based on the feature ranking process implemented by RF and SVM with linear kernel, the frequency that a patient shows a fever is associated with greater risk for having an unpleasant result. From the results, it is evident that our study is complementary to previous studies on treatment outcome classification. Based on the experiments and observations, it can be concluded that the clinical and medical datasets including treatment-related complications, would generally improve the treatment outcome prediction and classification for ALL patients.

There were shortcomings as well. The first one is the lack of an integrated information system. Paper-based records were a huge challenge

for the data gathering process, also the enormous volume of inpatient files, and only one person assigned to all of the data gathering processes made the work time-consuming and difficult, but worthwhile. Future works from the authors' perspective could be the analysis of treatment outcome of ALL patients utilizing the combination of different datasets, including gene expression data, image datasets and clinical data.

Ethics approval

The data used for this study has been rewarded the research ethics certificate with approval ID of IR.MODARES.REC.1398.149 and were authorized by a team of experts from MAHAK hospital.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

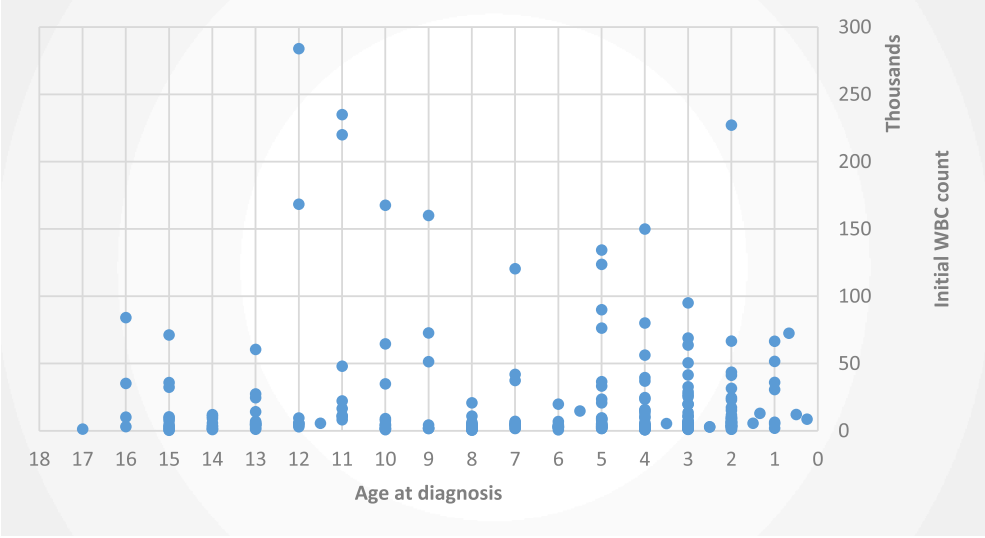
Acknowledgements

We would like to thank the staff members from the section of clinical hematology and oncology of MAHAK hospital. Besides, a special thanks to the MAHAK cancer research center for their assistance.

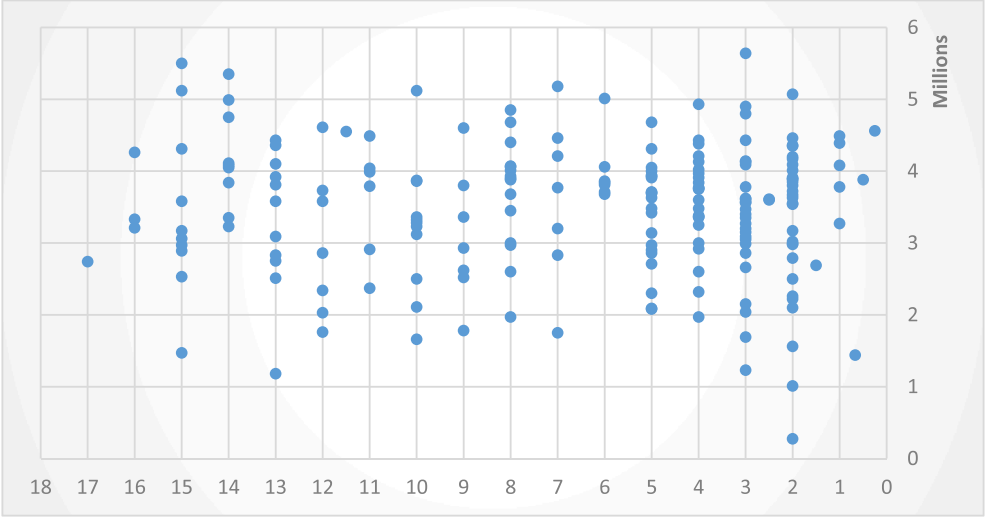
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2020.100399.

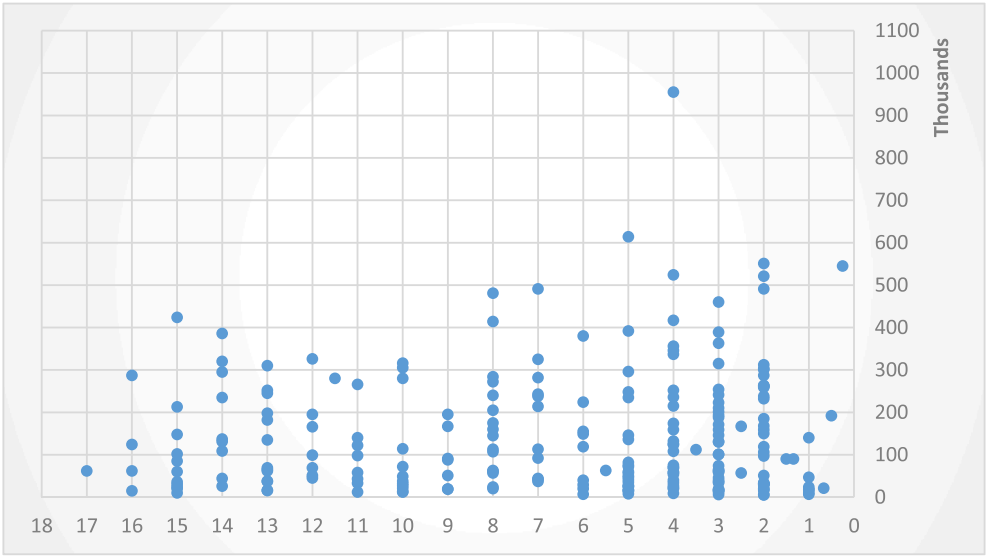
Appendix 1. Initial WBC count scaled on the age at diagnosis



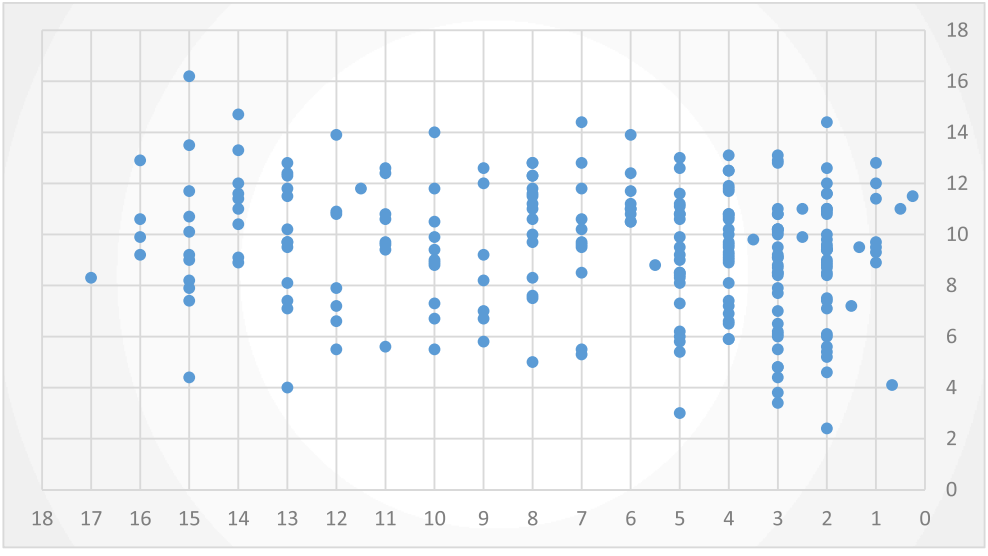
Appendix 2. Initial RBC count scaled on the age at diagnosis



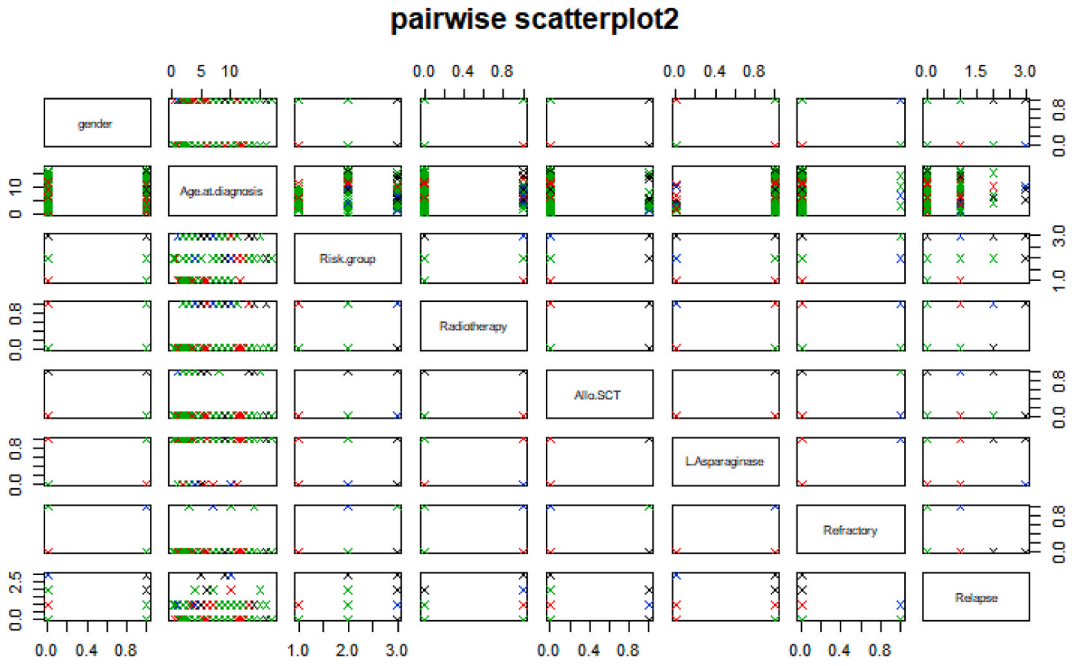
Appendix 3. Initial PLT count scaled on the age at diagnosis

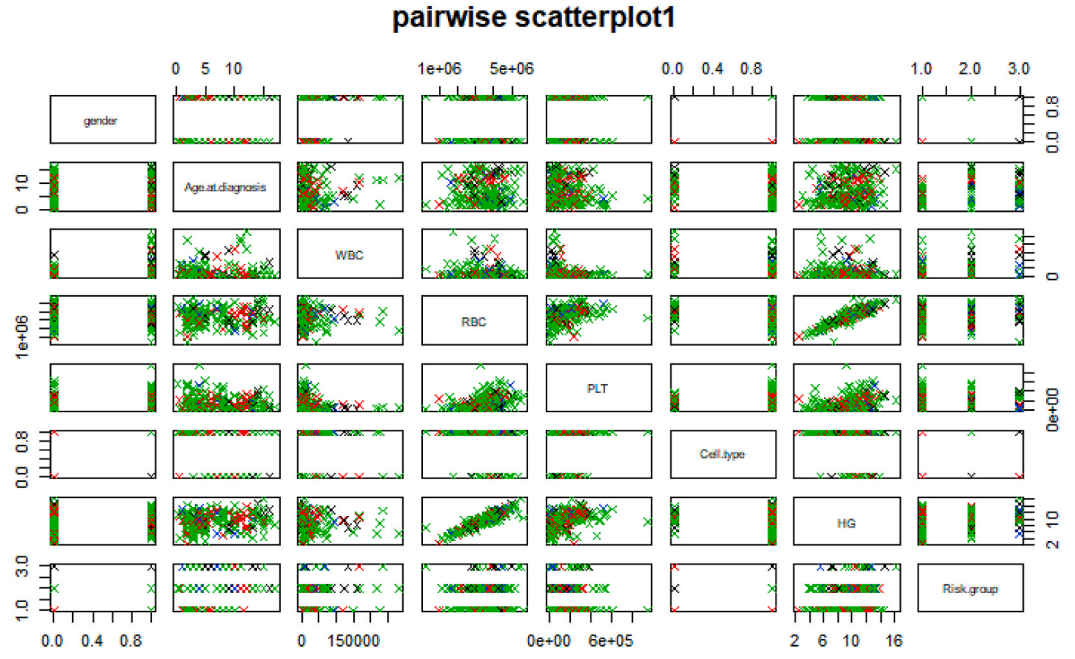


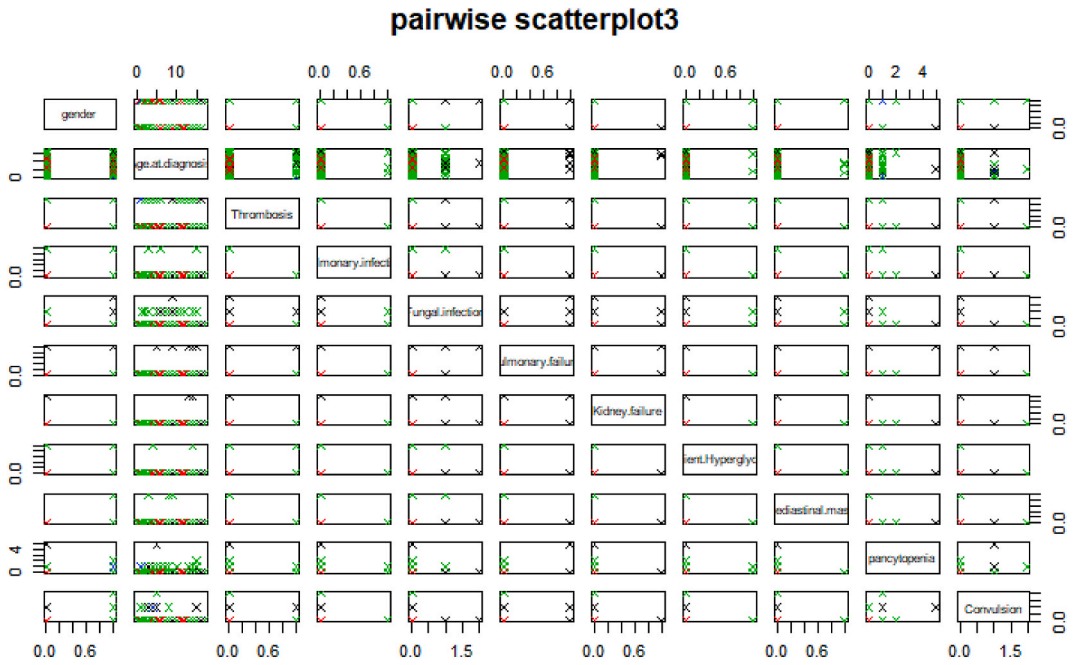
Appendix 4. Initial HG scaled on the age at diagnosis

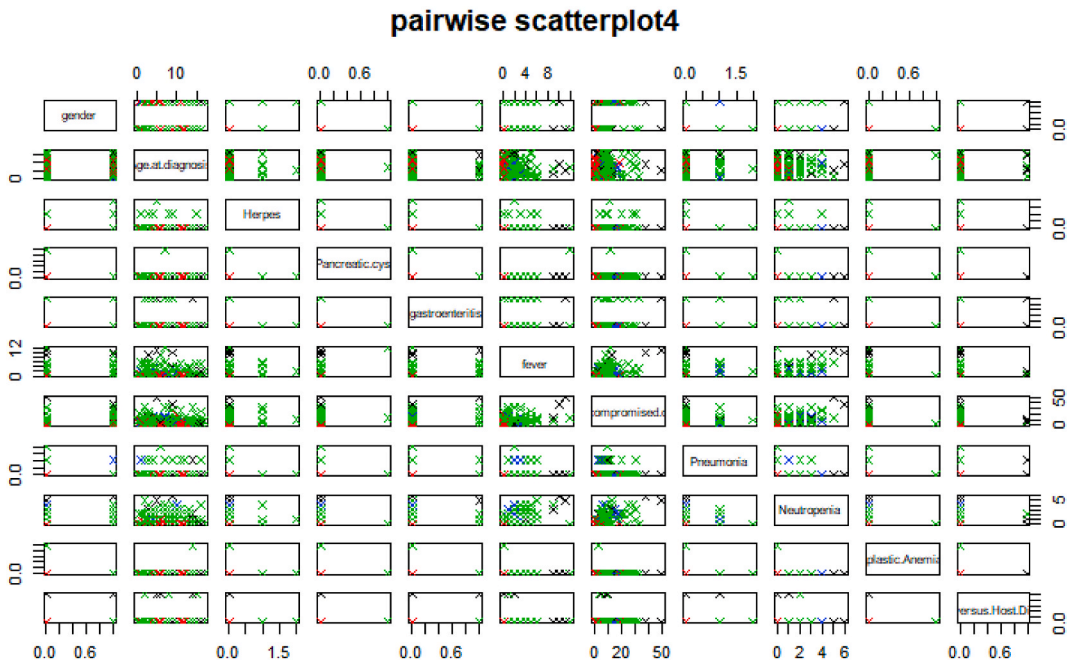


Appendix 5. The pairwise scatterplots of dataset variables









References

[1] Rohayanti H, Harun N, Yusoff M. Classification of blasts in acute leukemia blood samples using k-nearest neighbour. In: IEEE 8th International Colloquium on Signal Processing and its Applications. IEEE; 2012.

[2] Abd El-Nasser A, Shaheen M, El-Deeb H. Enhanced leukemia cancer classifier algorithm. In: Science and Information Conference. London, UK: IEEE; 2014.

[3] Alsalem MA, et al. A review of the automated detection and classification of acute leukaemia: coherent taxonomy, datasets, validation and performance measurements, motivation, open challenges and recommendations. Comput Methods Programs Biomed 2018;158:93–112.

[4] Mohapatra S, Patra D, Satpathi S. Image analysis of blood microscopic images for acute leukemia detection. In: International Conference on Industrial Electronics, Control and Robotics. Orissa, India: IEEE; 2010.

[5] Rawat J, et al. Review of leukocyte classification techniques for microscopic blood images. In: 2nd International Conference on Computing for Sustainable Global Development (INDIACom). New Delhi, India: IEEE; 2015.

[6] Belson M, Kingsley B, Holmes A. Risk factors for acute leukemia in children: a review. Environ Health Perspect 2007;115(1):138–45.

[7] Tran VN, et al. An automated method for the nuclei and cytoplasm of Acute Myeloid Leukemia detection in blood smear images. In: World Automation Congress (WAC). Rio Grande, Puerto Rico: IEEE; 2016.

[8] Laosai J, Chamnongthai K. Acute leukemia classification by using SVM and K-Means clustering. In: International Electrical Engineering Congress (iEECON). Chonburi, Thailand: IEEE; 2014.

[9] Labib NM, Malek MN. Data mining for cancer management in Egypt case study childhood acute lymphoblastic leukemia. ENFORMATIKA 2005;8:309–14.
[10] Berg SL, Poplack DG. Complications of leukemia. Pediatr Rev 1991;12(10):313–8.
[11] Wyatt KD, Bram RJ. Immunotherapy in pediatric B-cell acute lymphoblastic leukemia. Hum Immunol 2019;80(6):400–8. https://doi.org/10.1016/j.humimm.2019.01.011.
[12] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA A Cancer J Clin 2019;69(1):7–34. https://doi.org/10.3322/caac.21551.
[13] Aroop K, Nobuko H. Diagnosis and initial management of pediatric acute leukemia in the emergency department setting. Clin Pediatr Emerg Med 2018;19(2):135–44.
[14] Willenbrock H, et al. Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukemia using DNA microarrays. Leukemia 2004;18(7):1270–7.
[15] Morton G. Data mining the genetics of leukemia. In: School of computing. Kingston, Ontario, Canada: Queen's University; 2010.
[16] Yeoh EJ, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Canc Cell 2002;1(2):133–43.
[17] Pommet L, Margossian S, Burke M. Diagnosis and treatment-related complications of acute leukemia. 2019. p. 9–28.
[18] Veisani Y, Khazaei S, Delpisheh A. 5-year survival rate based on the types of leukemia in Iran: a meta-analysis. Caspian J Intern Med 2018;9(4):316–24.
[19] Bhojwani D, et al. Gene expression signatures predictive of early response and outcome in high-risk childhood acute lymphoblastic leukemia: a Children's Oncology Group Study. J Clin Oncol 2008;26(27):4376–84.
[20] Hosseinzadeh Kassani S, et al. A hybrid deep learning architecture for leukemic B-lymphoblast classification. In: International Conference on Information and Communication Technology Convergence (ICTC). IEEE; 2019.
[21] Ross ME, et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood 2003;102(8):2951–9.
[22] Singhal V, Singh P. Correlation based feature selection for diagnosis of acute lymphoblastic leukemia. In: WCI'15: Proceedings of the Third International Symposium on Women in Computing and Informatics; 2015. p. 5–9.
[23] Ongun G, et al. Feature extraction and classification of blood cells for an automated differential blood count system. In: IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222). Washington, DC, USA, USA: IEEE; 2001.
[24] Singh G, Bathla G, Kaur SP. A review to detect leukemia cancer in medical images. In: International Conference on Computing, Communication and Automation (ICCCA). Noida, India: IEEE; 2016.
[25] Banjar H, et al. Intelligent techniques using molecular data analysis in leukaemia: an opportunity for personalized medicine support system. Biomed Res Int 2017;2017:3587309.
[26] Shearer C. The CRISP-DM model: the new blueprint for data mining. J Data Warehousing 2000;(5):13–22.
[27] Mehrvar A, et al. Epidemiological features of childhood acute leukemia at MAHAK's pediatric cancer treatment and research center (MPCTRC), Tehran, Iran. Basic Clin Canc Res 2015;7(1):9–15.
[28] Rubnitz JE, Pui CH. Childhood acute lymphoblastic leukemia. Oncol 1997;2:374–80.
[29] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018;68(1):7–30.
[30] Siegel SE, et al. Pneumonia during therapy for childhood acute lymphoblastic leukemia. Am J Dis Child 1980;134:28–34.
[31] Caruso V, et al. Thrombotic complications in childhood acute lymphoblastic leukemia: a meta-analysis of 17 prospective studies comprising 1752 pediatric patients. Blood 2006;108(7):2216–22.
[32] Attarbaschi A, et al. Mediastinal mass in childhood T-cell acute lymphoblastic leukemia: significance and therapy response. Med Pediatr Oncol 2002;39(6):558–65.
[33] Oparaji JA, et al. Risk factors for asparaginase-associated pancreatitis: a systematic review. J Clin Gastroenterol 2017;51(10):907–13.
[34] Rowe JM. Graft-versus-disease effect following allogeneic transplantation for acute leukaemia. Best Pract Res Clin Haematol 2008;21(3):485–502.
[35] Zhou C, Jia Y, Motani M. Optimizing autoencoders for learning deep representations from health data. IEEE J Biomed Health Inform 2019;23(1):103–11.
[36] Beaulieu-Jones KB, Moore JM. Missing data imputation in the electronic health record using deeply learned autoencoders. Pac Symp Biocomput 2017;22:207–18.
[37] Stekhoven DJ, Buehlmann P. MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics 2012;28(1):112–8.
[38] Johnson RA, Wichern DW. Applied multivariate statistical analysis. 6th ed. Upper Saddle Rover, NJ: Prentice Hall; 2002.
[39] Ross KS, Page D. AUCμ: a performance metric for multi-class machine learning models. In: Proceedings of the 36th International Conference on Machine Learning, vol. 97. Long Beach, California: PMLR; 2019.