*Iran J Ped Hematol Oncol. 2026, Vol. 16, No. 2, 826-838*

Original Article

# Application of Machine Learning in the Prediction of Liver Iron Concentration Using T2* MRI in the Liver of Children with Acute Lymphoblastic Leukemia

Reza Sadeghikhoo[1] MD, Mohammadreza Bordbar[2] MD, Zahra Abedini[1] MD, Omid Reza Zekavat[2*]MD

[1] School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran.
[2] Hematology Research Center, Shiraz University of Medical Sciences, Shiraz, Iran.
*Corresponding Author: Dr. Omid Reza Zekavat, Professor of pediatric hematology and oncology, Hematology Research Center, Shiraz University of Medical Sciences, Shiraz, Iran. Email: Ozekavat@gmail.com. ORCID: https://orcid.org/0000-0002-0008-3198.

Received: October 28, 2025;
Accepted: March 10, 2025

## Abstract

**Background:** Machine Learning (ML) is a technique currently used to predict, diagnose, and treat diseases. Acute Lymphoblastic Leukemia (ALL) is the most common cancer among pediatric patients. A frequent complication in children with ALL is anemia, which leads to the need for recurrent blood transfusions. This, in turn, can result in increased Liver Iron Concentration (LIC). Currently, diagnosing LIC is performed using T2* MRI techniques; however, due to the limitations of MRI, employing ML techniques to predict LIC has become necessary.

**Materials and Methods:** In this retrospective cohort study, a collection of datasets was obtained from 66 children (mean age = 10.7 year) diagnosed with ALL, and three ML models were used, including Random Forest Classifier (RFC), Support Vector Classifier (SVC), and Logistic Regression (LR). Given the small sample size, a preprocessing step including feature standardization and SMOTE oversampling was taken only within the training datasets during cross-validation to prevent data leakage.

**Results:** Among the evaluated models, LR achieved the highest precision–recall (PR) Area Under the Curve (AUC) and receiver operating characteristic (ROC) AUC values (test PR AUC = 0.94, p-value = 0.002; CV PR AUC = 0.98 p-value < 0.001; test ROC AUC = 0.98, p-value = 0.002; CV ROC AUC = 0.98, p-value < 0.001). The permutation feature importance identified serum ferritin (SF) and transfusion volume per kilogram (TV/Kg) as the dominant predictors of LIC.

**Conclusion:** This study indicates that ML models are promising ones for predicting LIC. However, due to the limited sample size, future studies with larger cohorts are warranted to validate these findings.

***Keywords:*** Acute lymphoblastic leukemia, Liver iron, Machine learning, Magnetic resonance imaging





[ DOI: 10.18502/ijpho.v16i2.21345 ]

## Introduction

Machine Learning (ML) is a field within Artificial Intelligence (AI) that is increasingly applied in the diagnosis and management of diseases (1-3). It is implemented through various methods, two major subgroups of which are supervised and unsupervised methods (1, 4). Different techniques are used in supervised ML, such as Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) (1). Among supervised ML techniques, some are used for classification (e.g., RFC and SVC), whereas others are used for regression (e.g., LR) (5). Machine learning has broad applications in medicine, such as pediatric oncology, where ALL remains a major clinical challenge. Accordingly, the integration of AI into cancer care is increasingly important.

Acute Lymphoblastic Leukemia (ALL) is the most common pediatric cancer and the leading cause of cancer-related death among individuals younger than 20 years (6). The long-term survival rate for children with ALL is approximately 90 percent (7). Moreover, anemia is common in children with ALL and may result from both the underlying disease and chemotherapy (8). In this regard, blood transfusions are often required in pediatric patients. Liver Iron Concentration (LIC) from repeated blood transfusions is a frequent complication in these patients (8).

Excessive iron accumulation in the body can result in iron overload, affecting the heart, liver, endocrine system, and other organs (9). In the liver, iron deposition can cause oxidative stress, hepatocellular injury, fibrosis, and hepatocellular carcinoma (10). Therefore, early prevention, diagnosis, and treatment are crucial in this group of patients.

Liver biopsy has traditionally been regarded as the gold standard for the assessment of hepatic iron overload (11). Currently, Magnetic Resonance Imaging (MRI) with a T2* sequence is preferred for assessing hepatic and cardiac iron deposition (11, 12). However, the T2 technique has limitations. Children often cannot tolerate the duration of the MRI procedure and may require sedation or general anesthesia (13). Therefore, alternative approaches such as AI models are necessary for prediction and prevention.

Recent studies suggest that serum ferritin (SF) is a key marker for the early prediction of the disease, and prescribing iron chelation therapy can prevent future morbidity (14-19). Therefore, determining the optimal cutoff value remains challenging.

This study aims to evaluate the application of ML methods in predicting LIC in children with ALL. A secondary objective is to identify the most important factors contributing to the prediction of LIC. Notably, this study reports one of the early applications of ML methods, such as RFC and SVM, to predict LIC in children with ALL.

## Material and Methods

### *Patients and data*

A total of 66 pediatric patients with ALL, aged 5-18 years, were enrolled in this retrospective cohort study. There were 40 males and 26 females. The data collection was conducted following approval from the institutional ethics committee of Shiraz University of Medical Sciences (ethical code: IR.SUMS.MED.REC 1400.219). Written informed consent was obtained from the parents of all the participants.

Children were assessed during six months after the completion of treatment. In this six-month follow-up period, the participants received packed red blood cell transfusions as clinically indicated, and the cumulative transfusion volume was recorded and normalized to body weight (TV/Kg). After six months, T2* MRI was performed to assess LIC. An iron profile was established too, which included SF, serum iron (SI), and total iron-binding capacity (TIBC). The demographic data (age, sex, weight, and height) and the transfusion volumes were also recorded. The cumulative transfusion volume per kilogram of body weight (TV/kg) was determined. Table I presents the descriptive characteristics of the study cohort, including demographic variables, transfusion measures, and iron profile parameters. Additionally, T2* MRI data were obtained from the liver. The relaxation times reported in the MRI results were categorized into two classes: values indicating iron deposition (n = 21) and values indicating no iron deposition (n = 45). To



[ DOI: 10.18502/ijpho.v16i2.21345 ]

determine the hepatic iron overload, the relaxation times were classified according to a cutoff value; < 6.3 ms indicated iron deposition, whereas ≥ 6.3 ms was considered as no deposition (14).

Notably, no patients in this cohort received iron chelation therapy during the study period; based on the treating physician's judgment, any patient who required initiation of chelation during the follow-up period was excluded from the study.

### *Equipment*

MRI was conducted using a 1.5 Tesla Siemens Avanto machine for T2* MRI data. A Lenovo laptop featuring an Intel(R) Core (TM) i7-3612QM CPU was also used. The analyses in this project were based on Python within Visual Studio Code version 1.98.0 (user setup).

### *Data analysis and modeling*

In coding, the first step was data loading and preprocessing. The data were imported to Python from an Excel sheet that included a column for gender, which was categorical. There were six numerical columns for age, BMI, TV/Kg, TIBC, SI, and SF, and a binary column represented the T2* MRI results for the liver.

The data preprocessing included a few steps. First, to ensure equal scaling, all the numerical datasets were standardized (20). Second, the datasets were split into two categories; 80 percent of them were allocated for Cross-Validation (CV) to train and optimize the model parameters, and 20 percent were assigned to the evaluation of the performance of the models. Third, due to the imbalance in the binary data, the Synthetic Minority Over-Sampling Technique (SMOTE) (21) was used to train the data to oversample the minority classes. To prevent data leakage, all the preprocessing steps, including feature standardization and SMOTE oversampling, were implemented within a pipeline framework and applied exclusively to the training folds during the cross-validation. The independent test set was kept completely untouched; it was not involved in scaling parameter estimation or resampling procedures. Hyperparameter optimization was conducted using GridSearchCV integrated within this pipeline structure.

The next step involved training the models and optimizing the hyperparameters using the GridSearchCV tool (20). K-Fold CV (with K = 5) was completed to ensure the stability of the performance of the models.

The next step was the assessment of the models by computing the AUC for both the ROC and PR curves in the CV and test data; Additionally, for the test datasets, F-1 score, recall, precision, specificity, and accuracy were computed in all the models (22). To quantify uncertainty, 95% CIs for the ROC AUC and PR AUC was estimated through nonparametric bootstrapping with 1,000 resamples and replacement. For each bootstrap sample, the corresponding AUC and the 2.5th and 97.5th percentiles of the bootstrap distribution as the 95% CI were calculated. Any bootstrap resample that contained only one outcome class was excluded.

Another step was plotting and analyzing the permutation feature importance (22, 23) for the best model. The step next to it involved plotting an ROC curve (22, 24) to determine an optimized threshold for the SF, which was calculated using Youden's Index (25). Additionally, a PR curve was created based on the F1 score to derive the optimal threshold for SF.

To evaluate the score separation, the predicted probabilities were compared between the LIC-positive and LIC-negative groups using a one-sided Mann–Whitney U test. (26), and a one-sample t-test was used to assess whether the permutation importance scores significantly differed from zero (23).

### *Machine learning models*

Three supervised classification models including RFC, SVC, and LR were evaluated (4). First, RFC is an ensemble model that combines multiple decision trees and outputs the final class by aggregating the trees' predictions (27). It was included to capture potential non-linear relationships among the clinical variables. The second model was SVC. It is a supervised classifier that separates classes using a decision boundary and can model non-linear patterns through kernel functions (28). As for LR, it is a widely used and interpretable baseline model for binary outcomes that estimates the probability of a target class from input variables (29). All these models were tuned using the same CV and grid-search procedure described above.



[ DOI: 10.18502/ijpho.v16i2.21345 ]

### *Model evaluation*

Standard classification metrics including accuracy, precision, recall, F1-score, specificity, and AUC (ROC AUC and PR AUC) were used to evaluate the models (30). To monitor potential overfitting, the CV AUC were compared with the held-out test AUC values; larger discrepancies were interpreted as possible signs of limited generalizability in this dataset.

## Results

In this section, first, the results of the PR and ROC curves are presented along with the other standard metrics for evaluating the models. The permutation feature importance results for all the models are also presented to quantify the contribution of each clinical variable to LIC prediction. Finally, the ROC and PR curve results are provided to identify the optimal threshold for both SF and TV/Kg.

### *AUC of the PR and ROC curves*

Table II summarizes the PR-AUC results for the three classifiers in both cross-validation and the independent test set, along with 95% CIs and AUC gaps. Table III provides the corresponding ROC-AUC results.

For the RFC model, the AUC of the PR curve in the CV dataset was 0.91, with a 95% Confidence Interval (CI) of 0.25 to 1.00 and a P-value (P) of less than 0.001. Given the relatively small sample size, 95% CI and AUC values are emphasized as primary indicators of model performance, with p-values reported for reference only. The AUC of the PR curve for the test was 0.87 (95% CI: 0.85 to 1.00, P = 0.004), and the AUC gap of the CV and the test datasets was 0.07. As shown in Table II, the RFC achieved a test PR-AUC of 0.87 and a CV PR-AUC of 0.94, with an AUC gap of 0.07. In contrast the AUC of the ROC curve for the CV and test datasets was 0.95 (95% CI: 0.89 to 1.00, P < 0.001) and 0.95 (95% CI: 0.78 to 1.00, P = 0.004), respectively, reflecting a gap of 0.00. As shown in Table III, the RFC achieved a test ROC-AUC of 0.95 and a CV ROC-AUC of 0.95, indicating no performance gap. Figure 2 illustrates the PR and ROC curves of the RFC model.

In the SVC model, the AUC of the PR curve for the CV and test datasets was found to be 0.98 (95% CI: 0.94 to 1.00, P < 0.001) and 0.91 (95% CI: 0.51 to 1.00, P = 0.004), respectively, reflecting a gap of 0.07. As shown in Table II, the SVC model achieved a test PR-AUC of 0.91 and a CV PR-AUC of 0.98, corresponding to an AUC gap of 0.07. In contrast, the AUC of the ROC curve for the CV and test datasets was 0.98 (95% CI: 0.95 to 1.00, P < 0.001) and 0.95 (95% CI: 0.77 to 1.00, P = 0.004), respectively, indicating a gap of 0.03. As shown in Table III, the SVC model achieved a test ROC-AUC of 0.95 and a CV ROC-AUC of 0.98, with a gap of 0.03. Figure 2 displays the diagrams for the PR and ROC curves of the SVC model.

In the LR model, the AUC of the PR curve for the CV and test datasets was 0.98 (95% CI: 0.94 to 1.00, P < 0.001) and 0.94 (95% CI: 0.63 to 1.00, P = 0.002), respectively, which shows a gap of 0.03. Based on Table II, the LR model achieved the highest test PR-AUC of 0.94 and a CV PR-AUC of 0.98, with a minimal AUC gap of 0.03. Subsequently, the AUC of the ROC curve for the CV and test datasets was found to be 0.98 (95% CI: 0.95 to 1.00, P < 0.001) and 0.98 (95% CI: 0.85 to 1.00, P = 0.002), respectively, indicating a gap of 0.01. According to Table III, the LR model achieved a test ROC-AUC of 0.98 and a CV ROC-AUC of 0.98, indicating a negligible performance gap. Figure 2 illustrates the diagrams for the PR and ROC curves of the LR model.

The accuracy for the RFC model was 0.93, with a F1-score of 0.89, recall of 1.00, precision of 0.80, and specificity of 0.90. For the SVC model, the accuracy was 0.86, and the F1-score, recall, precision, and specificity were 0.80, 1.00, 0.67, 0.80, respectively. Finally, for the LR model, the accuracy was 0.93, with a F1-score of 0.89, recall of 1.00, precision of 0.80, and specificity of 0.90. For better visual differentiation, all the outputs of the models are displayed in Figure 3.

### *Permutation feature importance*

In the following analysis, the permutation feature importance is presented for all the models. The RFC indicates a feature importance of 76.3% for SF (P < 0.001), 15.6% for TV/Kg (P < 0.001), 4.1% for Age (P < 0.001), 1.9% for BMI (P = 0.002), 1.6% for TIBC (P = 0.999), 0.6% for SI (P = 0.780), and 0.0% for gender (P = 1.000). The SVC model has the following importance levels:



[ DOI: 10.18502/ijpho.v16i2.21345 ]

SF: 46.0% (P < 0.001), TV/Kg: 22.5% (P < 0.001), TIBC: 11.3% (P < 0.001), BMI: 9.1% (P < 0.001), Age: 7.0% (P < 0.001), SI: 2.2% (P = 0.086), and gender: 1.9% (P = 0.079). Finally, the LR model shows the following importance levels: SF: 82.1% (P < 0.001) and TV/Kg: 17.9% (P < 0.001), while TIBC, BMI, Age, SI, and gender all have an importance of 0.00% (P = 1.000). For better visual comparison, the permutation feature importance results of the models are illustrated in Figure 4.

### *PR and ROC curves for SF and TV/Kg*

The PR and ROC curve results for determining the optimal thresholds of SF and TV/kg are presented here. The optimal threshold for SF was 400.00 ng/ml, with a recall of 0.90 and a precision of 0.90. The PR AUC was 0.93 (95% CI: 0.83 to 1.00, P < 0.001), with a sensitivity of 0.90 and a specificity of 0.96, resulting in an ROC AUC of 0.97 (95% CI: 0.93 to 1.00, P < 0.001). The optimal threshold for TV/Kg was 29.79 ml/Kg, achieving a recall of 0.81 and a precision of 0.65. The PR AUC was 0.78 (95% CI: 0.61 to 0.91, P < 0.001), with a sensitivity of 0.81 and a specificity of 0.80, resulting in an ROC AUC of 0.87 (95% CI: 0.77 to 0.94, P < 0.001). The PR and ROC curves for determining the optimal thresholds of the SF and TV/Kg are provided in Figure 5.

*Table I. Descriptive data for the patients participating in this study*

| Variables | N | Range | Minimum | Maximum | Mean ± SD |
|---|---|---|---|---|---|
| **Age (year)** | 66 | 13 | 5 | 18 | 10.74 ± 4.42 |
| **Wt. (Kg)** | 66 | 83 | 13 | 96 | 39.32 ± 19.97 |
| **Ht. (m)** | 66 | 0.87 | 0.95 | 1.82 | 1.40 ± 0.24 |
| **BMI (Kg/m$^2$)** | 66 | 20.42 | 12.80 | 33.22 | 18.78 ± 4.36 |
| **Tx. N (n)** | 66 | 18 | 2 | 20 | 4.82 ± 4.28 |
| **TV (ml)** | 66 | 6810 | 210 | 7020 | 1423.64 ± 1577.00 |
| **TV/Kg(ml/Kg)** | 66 | 114.16 | 9.00 | 123.16 | 34.14 ± 30.22 |
| **TIBC (µg/dl)** | 66 | 202 | 225 | 427 | 314.82 ± 40.23 |
| **SI (µg/dl)** | 66 | 181 | 19 | 200 | 89.00 ± 33.50 |
| **SF (ng/ml)** | 66 | 3044 | 5 | 3049 | 401.47 ± 512.03 |

N: Number, Wt.: Weight, Ht.: Height, BMI: Body Mass Index, Tx. N: Transfusion Number, TV: Transfusion Volume, TV/Kg: Transfusion blood Volume per Kilogram body weight, TIBC: Total Iron Binding Capacity, SI: Serum Iron, and SF: Serum Ferritin.

*Table II. Comparison of outputs in different models based on the AUC of CV and Test datasets for the PR curve*

| | Test AUC | Test AUC 95 % CI | CV AUC | CV AUC 95 % CI | AUC gap | p-value test | p-value CV |
|---|---|---|---|---|---|---|---|
| **RFC** | 0.87 | 0.25-1.00 | 0.94 | 0.85-1.00 | 0.07 | 0.004 | < 0.001 |
| **SVC** | 0.91 | 0.51-1.00 | 0.98 | 0.94-1.00 | 0.07 | 0.004 | < 0.001 |
| **LR** | 0.94 | 0.63-1.00 | 0.98 | 0.94-1.00 | 0.03 | 0.002 | < 0.001 |

RFC: Random Forest Classifier, SVC: Supportive Vector Classifier, LR: Logistic Regression, CV: Cross Validation, AUC: Area Under the Curve, CI: Confidence Interval, PR: Precision-Recall, CV: Cross Validation



[ DOI: 10.18502/ijpho.v16i2.21345 ]

*Table III. Comparison of outputs in different models based on the AUC of CV and Test datasets for the ROC curve*

| | Test AUC | Test AUC 95 % CI | CV AUC | CV AUC 95 % CI | AUC gap | p-value test | p-value CV |
|---|---|---|---|---|---|---|---|
| **RFC** | 0.95 | 0.78-1.00 | 0.95 | 0.89-1.00 | 0 | 0.004 | 0.95 |
| **SVC** | 0.95 | 0.77-1.00 | 0.98 | 0.95-1.00 | 0.03 | 0.004 | 0.95 |
| **LR** | 0.98 | 0.85-1.00 | 0.98 | 0.95-1.00 | 0.01 | 0.002 | 0.98 |

RFC: Random Forest Classifier, SVC: Supportive Vector Classifier, LR: Logistic Regression, CV: Cross Validation, AUC: Area Under the Curve, CI: Confidence Interval, ROC: Receiver-Operating Characteristic, CV: Cross Validation

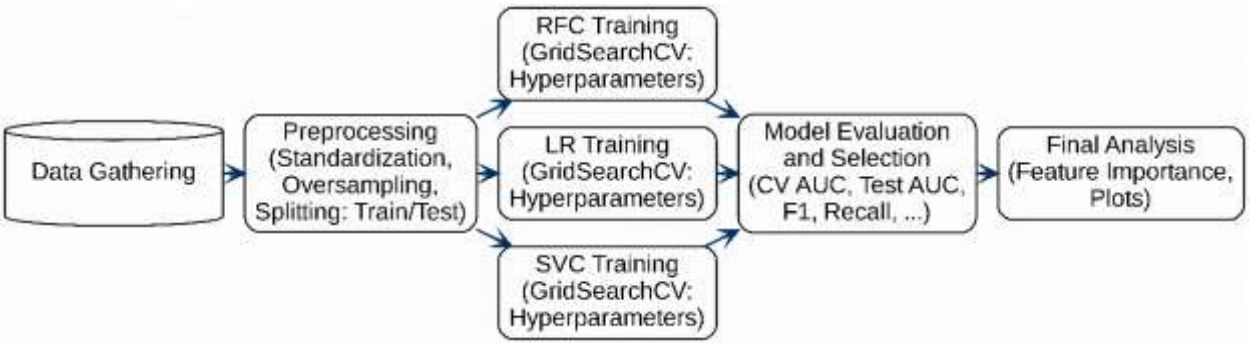


Figure 1. Overview of the steps in this study, from data collection to final analysis. AUC: Area Under the Curve; CV: Cross-Validation; F1: F1 Score; GridSearchCV: Grid Search Cross-Validation; LR: Logistic Regression; RFC: Random Forest Classifier; SVC: Support Vector Classifier



[ DOI: 10.18502/ijpho.v16i2.21345 ]

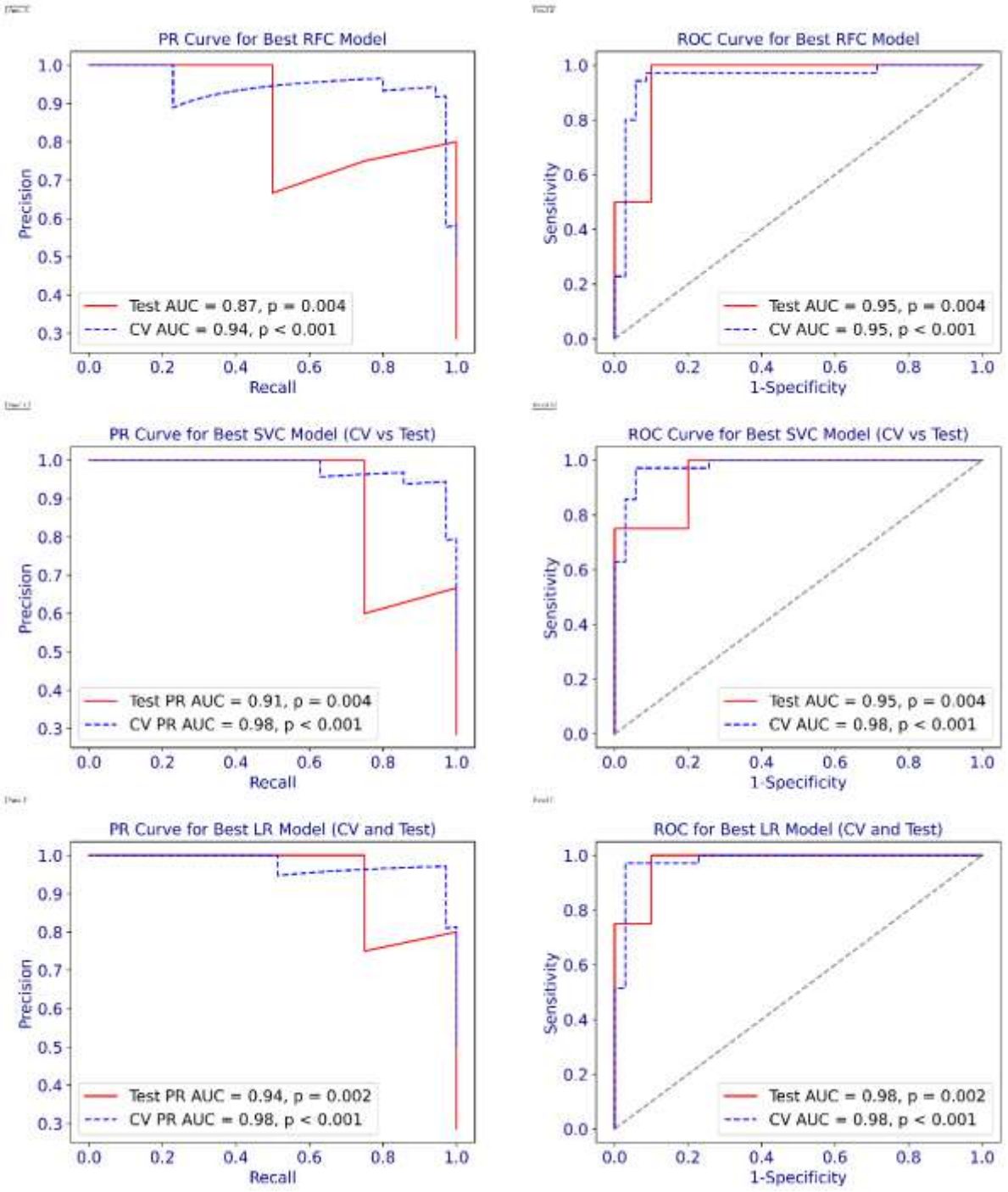


Figure 2. Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves for the Random Forest Classifier (RFC) model (A, B), Support Vector Classifier (SVC) model (C, D), and Logistic Regression (LR) model (E, F). AUC: Area Under the Curve; PR: Precision-Recall; ROC: Receiver Operating Characteristic; CV: Cross-Validation



[ DOI: 10.18502/ijpho.v16i2.21345 ]

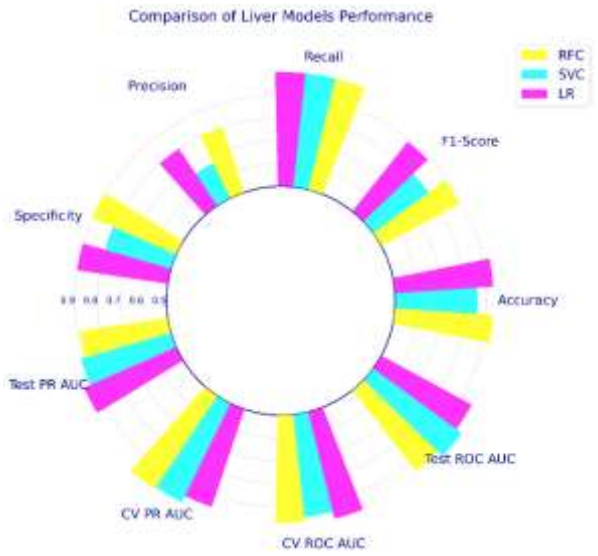


Figure 3. Visual comparison of model outputs across different classifiers. RFC: Random Forest Classifier; SVC: Support Vector Classifier; LR: Logistic Regression; CV: Cross-Validation; AUC: Area Under the Curve; PR: Precision–Recall; ROC: Receiver Operating Characteristic

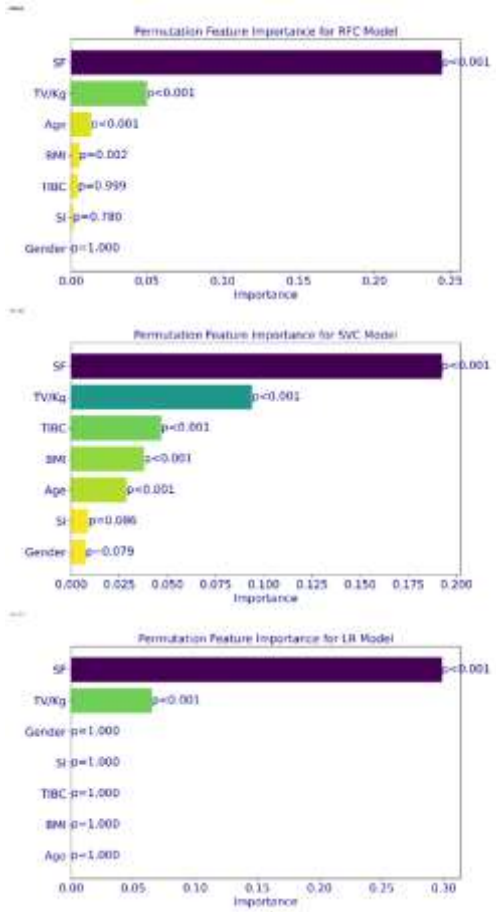


Figure 4. Permutation feature importance for all the models. RFC: Random Forest Classifier; SVC: Support Vector Classifier; LR: Logistic Regression; SF: Serum Ferritin; TV/kg: Transfusion Volume per Kilogram Body Weight; BMI: Body Mass Index; TIBC: Total Iron-Binding Capacity; SI: Serum Iron



[ DOI: 10.18502/ijpho.v16i2.21345 ]

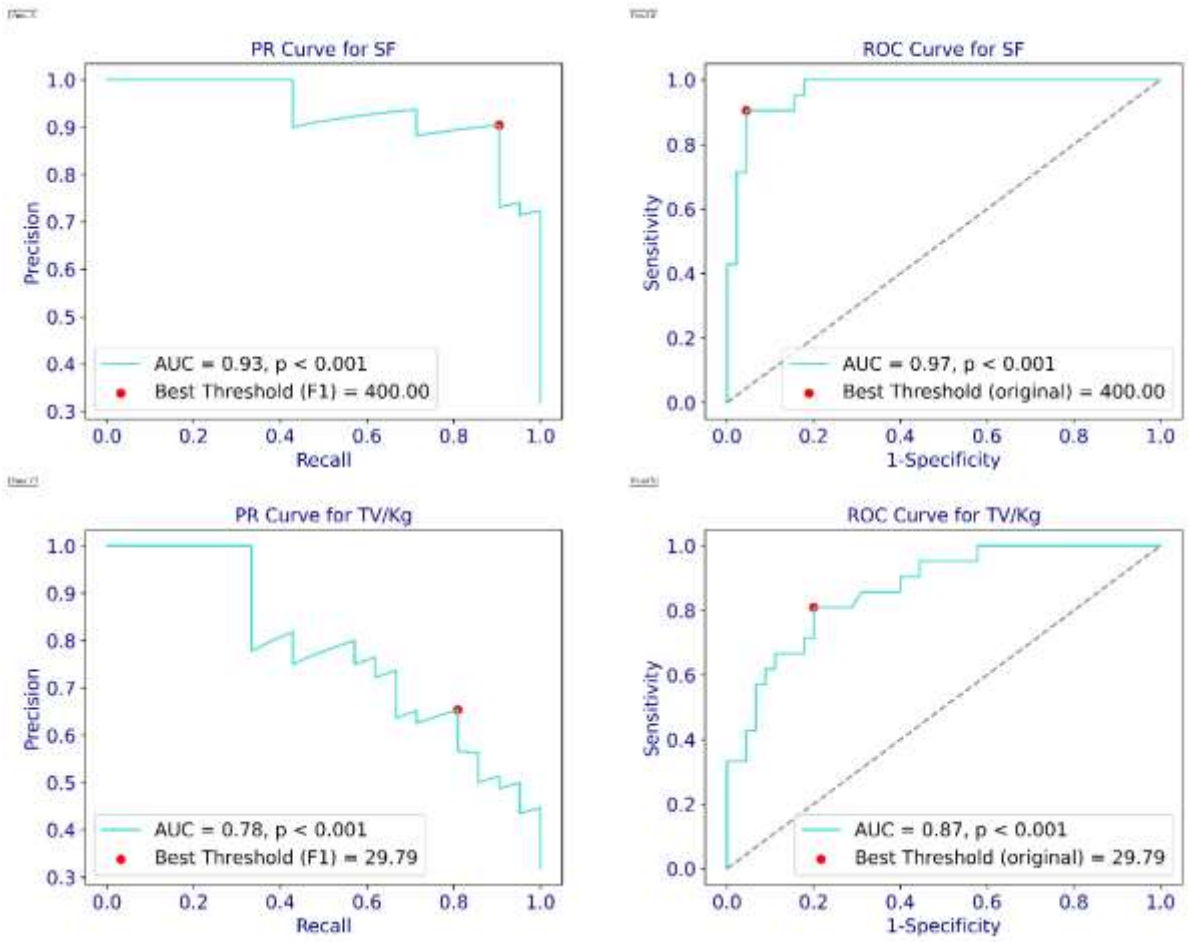


Figure 5. Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves used to determine thresholds for serum ferritin (SF) (A, B) and transfusion volume per kilogram body weight (TV/kg) (C, D) to distinguish between iron deposition and no-iron deposition. AUC: Area Under the Curve; PR: Precision-Recall; ROC: Receiver Operating Characteristic.

## Discussion

The models evaluated in this study demonstrated strong discrimination for predicting LIC in children with ALL, as indicated by the high AUC values (test PR AUC = 0.94; CV PR AUC = 0.98; test ROC AUC = 0.98; and CV ROC AUC = 0.98). The small differences in performance between the CV and test sets further support the robustness of these models. Among the various models tested, LR achieved the highest AUC, outperforming both RFC and SVC. The permutation feature importance consistently identified SF and TV/kg as the most influential predictors in all the models.

Given the class imbalance in the dataset (LIC present: n = 21; LIC absent: n = 45), the evaluation of the models prioritized recall and metrics based on the PR curve. A recall rate of 1.00 was achieved for all the models. The precision of SVC was 0.67, indicating a higher false positive rate. The primary objective of this study was to maximize the recall rate, making the reduced precision less critical in this context. The high PR AUC values across the models support their effectiveness in identifying positive cases. In the existing literature, previous studies have often focused primarily on ROC curves; however, this analysis produced strong ROC-based results, indicating effective class separation. Despite the high AUC values, the wide CI highlights significant uncertainty due to the small sample size. This emphasizes that interpretation prioritizes CI rather than statistical significance alone.

Few studies have applied ML approaches to evaluate LIC. A recent systematic review and meta-analysis suggests that ML and deep learning methods can enhance MRI-based LIC



[ DOI: 10.18502/ijpho.v16i2.21345 ]

quantification and classification of liver iron overload (31). Another study by Munikoty et al. (32) utilized multiple linear regression to assess iron overload, whereas many reports have relied primarily on descriptive statistics (33-35). The present study reports an early application of advanced ML methods, including RFC and SVC, to investigate LIC in children.

In this study, LR outperformed more complex models like RFC, probably due to the limited sample size and a predictor set consisting of a small number of routinely measured variables. In such settings, simpler linear models may be generalized better and be less susceptible to variance and overfitting, effectively capturing the dominant signal without being influenced by noise (36).

From a clinical perspective, once validated, the proposed approach could serve as a supportive tool by utilizing routinely available variables (e.g., SF and TV/kg) to estimate an individual patient's risk of LIC. Patients classified as higher risk could be prioritized for closer monitoring and earlier evaluation for chelation therapy, particularly in contexts with limited access to advanced imaging. However, due to the limited sample size and lack of external validation, the current model should not be used for clinical decision-making until confirmed in larger cohorts with independent validation.

SF has consistently been identified as a key indicator of iron overload. Nashwan et al. (18) reported a median SF level of 687.25 ng/ml in cases of iron overload. Alkindi et al. (33) found a strong correlation between SF and LIC but did not establish a specific cutoff. Similarly, Acar et al. (34) reported SF as an important marker. Sawicka-Zukowska et al. (8) noted a significant association between SF and LIC, while Munikoty et al. (32) suggested that an SF level exceeding 600 ng/ml may indicate iron overload. Alternatively, Sherief et al. (35) proposed a threshold of 205 ng/ml. The findings from this study support SF as a critical predictor of LIC; an exploratory threshold of 400 ng/ml was identified in this respect. TV/kg was also recognized as a significant predictor, with previous studies proposing cutoffs for transfusion exposure. Sawicka-Zukowska et al. (8) reported a cutoff of 100 ml/kg for assessing iron overload, whereas the present study identified a lower exploratory threshold of approximately 30 ml/kg over a six-month period. Additionally, age has been cited as a relevant factor (8), which is also supported by this analysis.

This study has several limitations. First, the small sample size (n = 66), which includes only 21 positive cases, may affect the stability of the model and increase the risk of overfitting. While CV is utilized, applying SMOTE to a small dataset can magnify noise and result in overly optimistic performance estimates. Therefore, high AUC values should be interpreted with caution. Second, the proposed cutoffs for SF at 400 ng/ml and TV/kg at approximately 30 ml/kg are exploratory and specific to this cohort; they require validation in larger independent studies before any clinical application. Third, the lack of external validation restricts the generalizability of the findings. Fourth, MRI assessments could not be conducted in children younger than 10 years, limiting the applicability of the results to younger patients. Fifth, due to the small sample size, the model inputs were restricted to a limited set of routinely available and clinically relevant variables (age, BMI, gender, SI, TIBC, SF, and TV/kg). Additional clinical or biochemical parameters were not included. Furthermore, beyond scaling and binary encoding of gender, formal multicollinearity screening and systematic feature selection, besides clinical judgment, were not performed. These steps should be included in future studies with larger cohorts. Finally, transfusion duration was not evaluated because all the participants were assessed over a fixed six-month period. However, understanding cumulative transfusion exposure is important for grasping the dynamics of iron overload and progressive organ iron deposition (37).

## Conclusion

This study demonstrated that AI can accurately predict LIC. Three ML methods were employed in the analysis. Given the critical nature of diagnosing positive cases in medicine, both the PR curve and ROC curves were



[ DOI: 10.18502/ijpho.v16i2.21345 ]

assessed for all the models. The results from the PR curve indicated that all the models could effectively predict the positive cases, while the ROC curve demonstrated good discrimination between the positive and negative classes.

Among the models evaluated, the LR model was identified as the optimal one based on both ROC and PR AUC results, along with a slight gap between the test and CV AUCs. Further analysis of feature importance through the permutation methods revealed that SF was the most significant factor influencing LIC, with the second most important factor being TV/Kg. The ROC and PR curves for SF and TV/Kg indicated the optimal threshold of 400 ng/ml for SF, while it was 30 ml/Kg for TV/Kg over a period of six months. It is to be noted that the proposed thresholds should be considered exploratory, thus requiring external validation before any clinical implementation.

In addition, iron deposition can negatively impact children with ALL. Although diagnostic liver biopsies in children come with certain risks, T2* MRI also has its limitations. This study demonstrates that utilizing AI provides significant advantages by assisting physicians in diagnosing iron deposition and treating patients more effectively. However, due to the small sample size, the study has some limitations. Conducting future research with a larger sample size and employing deep learning techniques for more reliable results are recommended.

## Availability of Data

All data is available in manuscript.

## Ethical Considerations

The data collection was conducted under the ethics code IR.SUMS.MED.REC 1400.219 from the institutional ethics committee of Shiraz University of Medical Sciences. Also, written informed consent was obtained from the parents of all the participating children.

## Acknowledgements

The authors thank all those who participated in this research. They also confirm that no AI tools were used in the study design, data analysis, or generation of the scientific content of this manuscript.

## Authors' Contributions

Reza Sadeghikhoo (first author): Conceptualization, formal analysis, and writing of the original draft; Mohammadreza Bordbar: Data curation (MRI acquisition) and investigation; Zahra Abedini: Formal analysis and writing of the original draft; Omid Reza Zekavat (corresponding author): Data curation (MRI acquisition) and investigation. All the authors reviewed and approved the final version of the manuscript and agree to be accountable for all the aspects of the work.

## Funding

This study received no fund from public, commercial, or non-profit agencies.

## Conflict of Interest

The authors declare that they have no conflicts of interest.

## References

1. Mahesh B. Machine Learning Algorithms - A Review. Int J Sci Res 2020; 9(1): 381-386.

2. Shehab M, Abualigah L, Shambour Q, Abu-Hashem MA, Shambour MKY, Alsalibi AI, et al. Machine learning in medical applications: a review of state-of-the-art methods. Comput Biol Med 2022; 145: 105458.

3. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019; 19(1): 64.

4. Valkenborg D, Geubbelmans M, Rousseau AJ, Burzykowski T. Supervised learning. Am J Orthod Dentofacial Orthop 2023; 164(1): 146-149.

5. Gupta V, Mishra VK, Singhal P, Kumar A. An overview of supervised machine learning algorithm. In: Proceedings of the 2022 11th International Conference on System Modeling and Advancement in Research Trends (SMART). Moradabad, India: IEEE 2022: 87-



[ DOI: 10.18502/ijpho.v16i2.21345 ]

92.

6. Smith MA, Seibel NL, Altekruse SF, Ries LAG, Melbert DL, O'Leary M, et al. Outcomes for children and adolescents with cancer: challenges for the twenty-first century. J Clin Oncol 2010; 28(15): 2625-2634.

7. Neaga A, Jimbu L, Mesaros O, Bota M, Lazar D, Cainap S, et al. Why do children with acute lymphoblastic leukemia fare better than adults? Cancers 2021; 13(15): 3886-3889.

8. Sawicka-Zukowska M, Kretowska-Grunwald A, Kania A, Topczewska M, Niewinski H, Bany M, et al. Iron overload in children with acute lymphoblastic and acute myeloblastic leukemia—experience of one center. Cancers 2024; 16(2): 367.

9. Hsu CC, Senussi NH, Fertrin KY, Kowdley KV. Iron overload disorders. Hepatol Commun 2022; 6(8): 1842-1854.

10. Pietrangelo A. Iron and the liver. Liver Int 2016; 36 (Suppl 1): 116-123.

11. Sohal A, Kowdley KV. A review of new concepts in iron overload. Gastroenterol Hepatol 2024; 20(2): 98-107.

12. Mirbehbahani NB, Vaseghi G, Rashidbaghan A, Vakili A, Jahazi A. Comparison of magnetic resonance imaging T2 results in beta-thalassemia patients treated by deferasirox or combination of deferoxamine and deferiprone. Iran J Pediatr Hematol Oncol 2020; 10(4): 241-249.

13. Thestrup J, Hybschmann J, Madsen TW, Bork NE, Sørensen JL, Afshari A, et al. Nonpharmacological interventions to reduce sedation and general anesthesia in pediatric MRI: a meta-analysis. Hosp Pediatr 2023; 13(10): e301-e313.

14. Zekavat OR, Fallah Tafti F, Bordbar M, Parand S, Haghpanah S. Iron overload in children with leukemia: identification of a cutoff value for serum ferritin level. J Pediatr Hematol Oncol 2024; 46(2): e137-e142.

15. Sadighi S, Sahinoglu E, Kubba AH, Patel J, Hosseini F, Shafiee MA, et al. Impact of serum ferritin and iron overload on acute myeloid leukemia outcomes: a systematic review and meta-analysis. Asian Pac J Cancer Prev 2024; 25(9): 2951-2962.

16. Nashwan AJ, Yassin MA, Mohamed Ibrahim MI, Abdul Rahim HF, Shraim M. Iron overload in chronic kidney disease: less ferritin, more T2* MRI. Front Med 2022; 9: 865669.

17. Jäger L, Rachamin Y, Senn O, Burgstaller JM, Rosemann T, Markun S. Ferritin cutoffs and diagnosis of iron deficiency in primary care. JAMA Netw Open 2024; 7(8): e2425692-e242695.

18. Delibaş D, Evrimler Ş, Ercan K, Gümüş M, Sarıyıldırım A, Arslan H. Iron overload in hemodialysis patients: comparison of serum iron parameters with T2* MRI sequence. J Clin Ultrasound 2024; 52(2): 124-130.

19. Edalatkhah R, Kargar M, Yazdanparast M. Analysis of how serum ferritin and the aspartate aminotransferase-to-platelet ratio index (APRI) are correlated to hepatic MRI T2* findings in children with beta-thalassemia major. Iran J Pediatr Hematol Oncol 2024; 14(3): 188-195.

20. Garreta R, Moncecchi G. Learning scikit-learn: machine learning in Python. Birmingham: Packt Publishing; 2013;1-10.

21. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 2017; 18(17): 1-5.

22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011; 12: 2825-2830.

23. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010; 26(10): 1340-1347.

24. Hogan J, Adams NM. On averaging ROC curves. Trans Mach Learn Res 2023; 4: 123.

25. Hassanzad M, Hajian-Tilaki K. Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review. BMC Med Res Methodol 2024; 24(1): 84.

26. MacFarland TW, Yates JM. Mann–Whitney U test. In: MacFarland TW, Yates JM, editors. Introduction to Nonparametric Statistics for the Biological Sciences Using R. Cham: Springer; 2016; 103-132.

27. Becker T, Rousseau AJ, Geubbelmans M, Burzykowski T, Valkenborg D. Decision trees



[ DOI: 10.18502/ijpho.v16i2.21345 ]

and random forests. Am J Orthod Dentofacial Orthop 2023; 164(6): 894-897.

28. Birzhandi P, Kim KT, Lee B, Youn HY. Reduction of training data using parallel hyperplane for support vector machine. Appl Artif Intell 2019; 33(6): 497-516.

29. Nusinovici S, Tham YC, Chak Yan MY, Wei Ting DS, Li J, Sabanayagam C, et al. Logistic regression was as good as machine learning for predicting major chronic diseases. J Clin Epidemiol 2020; 122: 56-69.

30. Tharwat A. Classification assessment methods. Appl Comput Inform 2021; 17(1): 168-192.

31. Elhaie M, Koozari A, Alshammari QT. Machine and deep learning for MRI-based quantification of liver iron overload: a systematic review and meta-analysis. Radiologie (Heidelb) 2025; 65(Suppl):10-20.

32. Munikoty V, Singh SK, Anmol B, Prateek B, Savita VA, K RM, et al. Estimation of iron overload with T2 MRI in children treated for hematological malignancies. Pediatr Hematol Oncol 2023; 40(4): 315-325.

33. Alkindi S, Panjwani V, Al-Rahbi S, Al-Saidi K, Pathare AV. Iron overload in patients with heavily transfused sickle cell disease—correlation of serum ferritin with cardiac T2* MRI, liver T2* MRI, and R2-MRI (FerriScan). Front Med 2021; 8: 731102.

34. Acar S, Gözmen S, Bayraktaroğlu S, Acar SO, Tahta N, Aydınok Y, et al. Evaluation of liver iron content by magnetic resonance imaging in children with acute lymphoblastic leukemia after cessation of treatment. Turk J Haematol 2020; 37(4): 263-270.

35. Sherief LM, Beshir M, Saleem SN, Elmozy W, Elkalioubie M, Soliman BK, et al. Assessment of transfusion-induced iron overload with T2* MRI in survivors of childhood acute lymphoblastic leukemia: a case-control study. Hematol Transfus Cell Ther 2024; 46 (Suppl 6): S263-S271.

36. Mehta P, Wang CH, Day AGR, Richardson C, Bukov M, Fisher CK, et al. A high-bias, low-variance introduction to machine learning for physicists. Phys Rep 2019; 810: 1-124.

37. Coates TD. Iron overload in transfusion-dependent patients. Hematology Am Soc Hematol Educ Program 2019; (1): 337-344.



[ DOI: 10.18502/ijpho.v16i2.21345 ]