EURASIAN JOURNAL OF MATHEMATICAL AND COMPUTER APPLICATIONS
ISSN 2306–6172
Volume 12, Issue 4 (2024) 32 – 46

PREDICTING HOSPITAL PATIENT READMISSION BY ANALYZING ELECTRONIC HEALTH RECORD WITH INTERPRETABLE MACHINE LEARNING

Bunyamin H., Wargasetia T. L., Kasih J.

Abstract Hospital patient readmission is defined as a situation where a patient is treated again in a hospital after she is discharged within a specific time frame: 30 days, for example. This research aims to predict whether or not a patient will be readmitted from a hospital by applying predictive modeling which is learned from historical data. Our patient dataset is extracted from MIMIC-IV, which consists of an electronic health record dataset in Beth Israel Deaconess Medical Center (BIDMC) from year 2008 to 2019. Our experiments utilize four categories of models that are linear (logistic regression and linear discriminant analysis), non-linear (K-nearest neighbors, naïve Bayes, decision tree, and support vector machines), ensemble (bagging classifier, random forests, and extra trees), and boosting models (adaboost, stochastic gradient boosting). The performance evaluation of each model is using balanced accuracy because of imbalanced classes in our dataset. Additionally, each model is processed through 10-fold cross-validation and followed by a hyperparameter tuning process which eventually reports that the tree-based models, such as decision trees, extra trees, and random forests achieve the highest balanced accuracy. This study also identifies the features that significantly influenced the model's predictions by utilizing the cumulative reduction in both the mean and standard deviation of impurity and two global model-agnostic techniques, that are permutation feature importance (PFI) and SHapley Additive exPlanations (SHAP). The results obtained from these three different approaches are consistent, highlighting that the average levels of hematocrit, sodium, and platelets in the blood, coupled with the duration between a patient's registration and discharge from the hospital are critical features that have a substantial impact on the prediction outcomes.

Keywords: Hospital patient readmission, MIMIC-IV, Machine learning, Random forests, Mean-impurity-decreased-based features, Permutation Feature Importance, SHAP.

AMS Mathematics Subject Classification: 68T05, 68T07.

DOI: 10.32523/2306-6172-2024-12-4-32-46

1 Introduction

A hospital patient readmission is defined as a situation where a patient whom a hospital has discharged comes again to the hospital after a certain period, for example, 1 month or 3 months [1]. Mostly, a readmission infers that a patient has not fully recovered; consequently, the readmission even more makes a patient's treatment more costly than the absence of a readmission. Besides being a reliable indicator of the effectiveness and quality of hospital treatment provided to patients, readmission rates have been used as a publicly reported metric for comparing hospitals and determining reimbursement of

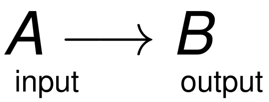


Figure 1: Supervised learning is identical to learn a learning between A and B.

hospital services [25, 11, 18, 16, 20]. Moreover, reducing the number of hospital readmissions can significantly decrease financial and healthcare redundancies; consequently, the quality of care is improved [8, 4].

Nowadays, most attention to reducing the amount of hospital readmissions shifts to utilizing Machine Learning (ML) as predictive models. Additionally, several data sources from Electronic Health Record (EHR) systems and Healthcare Information Systems (HIS) have been released publicly and can be used to train the models [27, 20, 6]. The type of model training which has been widely adopted and is so successful for many applications is supervised learning. Supervised learning treats the model as a function and enables the function to learn the relationship between input (A) and output (B) through the dataset as depicted in Fig. 1 [22, 13, 24].

A wide range of supervised learning algorithms and features have been proposed for readmission prediction and compiled into several surveys [27, 6, 12, 26]. Wang et al. [27] summarizes that the predominant focus of research involves utilizing regression-based approaches (specifically logistic regression), neural networks, and ensemble techniques like bagging, boosting, random forest, and gradient boosting, among the available methods. Our work utilizes similar learning algorithms and adds Light Gradient Boosting Machine (LightGBM) [15] to our list of models. Furthermore, they classify features into demographic, admission and discharge information, clinical information, hospital information, textual information, hybrid information, and latent features. Our research uses demographic, admission and discharge information, and clinical information features from the MIMIC-IV dataset [14].

Chen et al. [6] report that Random Forest (RF) is the mostly used learning algorithm in their study, followed by Gradient Boosting Machine (GBM), neural network-based algorithms, and Support Vector Machines (SVM). Particularly, GBM outperforms the other algorithms in terms of Receiver Operating Characteristic Area Under Curve (ROC AUC). Another effective learning algorithm is neural networks (NN). Deep Neural Networks (DNN) achieved a higher ROC AUC compared to Logistic Regression (LR), Random Forest, and Naïve Bayes (NB), displaying superior performance. Additionally, this article describes identified factors related to heart failure (HF) such as "age", Charlson comorbidity index, number of admissions in 6 or 12 months before index HF admission, and drug use before HF admission. Our work also compares all the mentioned algorithms; furthermore, we utilize factors that are far more comprehensive and general for any disease than their specific features.

Huang et al. [12] summarize that tree-based methods, Neural Networks (NN), logistic regression with regularization, and Support Vector Machines (SVM) emerge as the prevalent algorithms. They also outline the percentages of methods for model validation across studies, that are internal validation (77%), training/testing split (49%), resampling (28%), external validation (9%), and no validation (6%). Our study also includes a comparison of all the algorithms that have been mentioned and additional algorithms; therefore, we employ a greater number of algorithms, making our list more extensive than theirs.

Teo et al. [26] provide a summary indicating that widespread algorithms include LR, SVM, Decision Tree (DT) and its variations, such as RF and GBM. Among the models based on neural networks, DNNs or Multilayer Perceptrons (MLPs) have gained extensive usage. Furthermore, they argue that an enormous volume of electronic clinical data does not guarantee improving predictive ability because of the absence of relevant data. However, while certain elements, like social factors, have been demonstrated to be linked to a heightened risk of readmission, this information is not easily accessible within healthcare institutions.

de Sá et al. [7] also describe the development of a predictive machine learning algorithm for patient readmission, utilizing data from the MIMIC IV dataset. However, they extract fewer features (14) than ours (27). Specifically, prominent features (number of mean hematocrit, number of mean plaletet, number of mean sodium, and discharge duration) which contribute significantly to predictions are not extracted. Therefore, our ROC AUC is higher than theirs.

Assaf et al. [2] conduct a study comparable to ours, but they employed the MIMIC III dataset, a predecessor to MIMIC IV. Despite using a different dataset, their findings align with ours in that the Random Forest classifier demonstrates the highest accuracy. However, their study focuses solely on prediction outcomes, while our research goes further by identifying the key features that influence the prediction results.

This research aims to employ machine learning to predict of readmission occurrence. In particular, the MIMIC-IV dataset [14] is utilized in this study. As far as we know, this research represents the initial endeavor to develop general-purpose hospital readmission prediction models without concentrating on particular diseases. A comparative analysis was conducted to determine the optimal classification accuracy, employing eleven machine learning algorithms, namely Logistic Regression (LR), Linear Discriminant Analysis (LDA), k-Nearest Neighbors (kNN), Naïve Bayes (NB) [3], Decision Tree (DT), Support Vector Machines (SVM), Bagging Classifier (BC), Random Forest (RF), Extra Trees (ET), Adaboost (AB), and Stochastic Gradient Boosting (SGB). Additionally, we also investigate the most crucial features that have the most significant impact on predicting hospital readmissions.

2 Datasets and Methods

2.1 Datasets

This research utilizes MIMIC-IV, a dataset of electronic health records covering admissions from 2008 to 2019 at Beth Israel Deaconess Medical Center (BIDMC) [14]. MIMIC-IV consists of three modules, namely `hosp`, `icu`, and `note`. The `hosp` module contains records of patient admissions, discharges, and transfers. The `icu` module records all information documented in the Intensive Care Unit (ICU) and the `note` module consists of summaries of discharge information and radiology reports.

The features used in this study are taken and extracted from the `hosp` and `icu` modules. Particularly, there are two types of features in the dataset. The first type is patients' numerical information, which are easily extracted, such as heart rate mean, standard deviation (std) of diastolic blood pressure, std of respiration rate, calcium mean, potassium mean, std of calcium, std of potassium, diastolic blood pressure mean, respiration rate mean, std of glucose, std of systolic blood pressure, hematocrit mean, sodium mean, std of hematocrit, std of sodium, glucose mean, systolic blood pressure mean, std of heart rate, albumin mean, platelet count mean, std of albumin, and std of platelet count. The second type of features are features that must be constructed, such as age, length of stay, discharge duration, number of transfers, and Charlson comorbidity index.

2.2 Methods

The first step in this research methodology is the creation of features. First, the creation of the age feature is described. The patient's age is not explicitly given in the dataset to protect the patient's personal information. The information provided is the anchor age which is not the actual age, the anchor year which is the reference year, and the registration time of the patient. Therefore, the patient's age can be calculated using the following formula:

$$\text{age} = \text{admission\_time} - \text{anchor\_year} + \text{anchor\_age}.$$

Next is the discussion of ethnicity features. The ethnicity feature is already present in the dataset, but due to the large number of ethnicities, this feature is subjected to normalization by focusing on 4 (four) major ethnic types such as white, Latina, Asian, and black. Ethnicities other than these four categories are categorized as others. The discharge location feature (where the patient has finished undergoing treatment or medication) has 3 values, that are home, medical facility, and others. Furthermore, the discharge duration feature is computed by calculating the time difference from registration in the hospital and discharge from the same hospital. The readmission feature (is readmission or not?) is created by calculating the time when a patient re-registers. If the time exceeds 30 days, the patient is categorized as a readmission case. The number of transfers feature is calculated based on the number of times a patient moves from regular inpatient care to ICU care. We then calculate the amount of time a patient stays in the ICU, also known as the length of stay (LOS) [1]. The Charlson comorbidity index feature is generally calculated by means of

$$\begin{aligned}\text{Charlson Comorbidity Index} = \text{age conversion} + &\sum_{\text{disease}\in A} I(\text{disease}) + \\ &\max\{I(\text{mild liver disease}), 3 \times I(\text{severe liver disease}\}) + \\ &\max\{2 \times I(\text{diabetes with cc}), I(\text{diabetes without cc})\} + \\ &\max\{I(\text{malignant cancer}), 6 \times I(\text{mst})\} + \\ &2 \times I(\text{paraplegia}) + 2 \times I(\text{renal disease}) + \\ &6 \times I(\text{AIDS}) \qquad (1)\end{aligned}$$

where $I(\text{disease}) = 1$ if there is disease in the patient and 0 if there is no disease in the patient, cc = chronic complication, and mst = metastatic solid tumor, and A = all

```
if age $\leq 50$ then
    conversion $\leftarrow 0$
else if age $\leq 60$ then
    conversion $\leftarrow 1$
else if age $\leq 70$ then
    conversion $\leftarrow 2$
else if age $\leq 80$ then
    conversion $\leftarrow 3$
else
    conversion $\leftarrow 4$
end if.
```

Figure 2: Pseudocode.

diseases except mild liver disease, severe liver disease, diabetes with cc, diabetes without cc, malignant cancer, mst, paraplegia, renal disease, and Acquired Immunodeficiency Syndrome (AIDS).

The diseases that are considered important in calculating Charlson comorbidity index are myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, rheumatic disease, peptic ulcer disease, and the diseases mentioned in Equation 1. The age conversion is calculated according to the following pseudocode (see Fig. 2).

To summarize, the total number of features used is 34, with 27 numerical features and 7 categorical features depicted in Tab. 1 and Tab. 2, respectively. Furthermore, these categorical features are converted into one-hot encoding forms [10].

Table 1: A list comprising 27 numerical features

Age	discharge duration	number of transfers
Length of stay	Mean Diastolic BP	Mean Glucose
Mean Heart Rate	Mean Resp Rate	Mean Systolic BP
Std Diastolic BP	Std Glucose	Std Heart Rate
Std Resp Rate	Std Systolic BP	Mean Albumin
Mean Calcium	Mean Hematocrit	Mean Platelet Count
Mean Potassium	Mean Sodium	Std Albumin
Std Calcium	Std Hematocrit	Std Platelet Count
Std Potassium	Std Sodium	Charlson comorbidity index

Table 2: A list comprising 7 categorical features

Admission type	Discharge location	Insurance
Race	Gender	First care unit
	Last care unit	

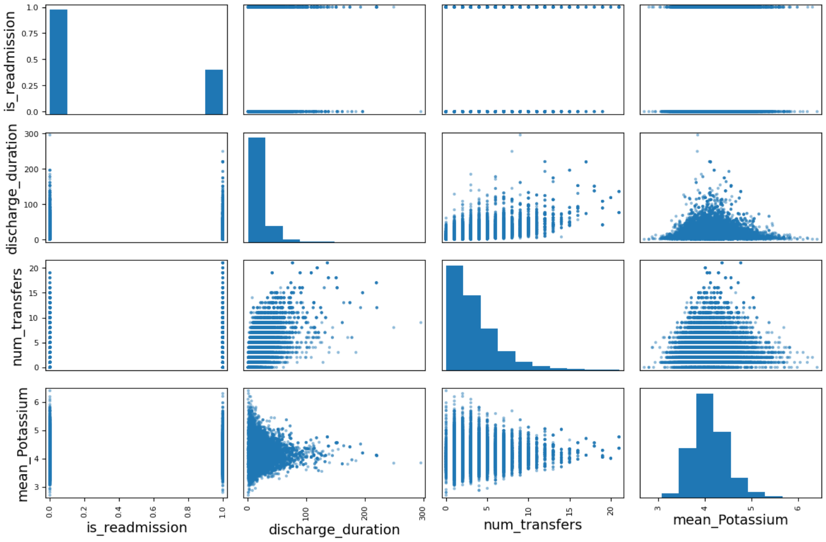


Figure 3: Scatter plot between the numerical features.

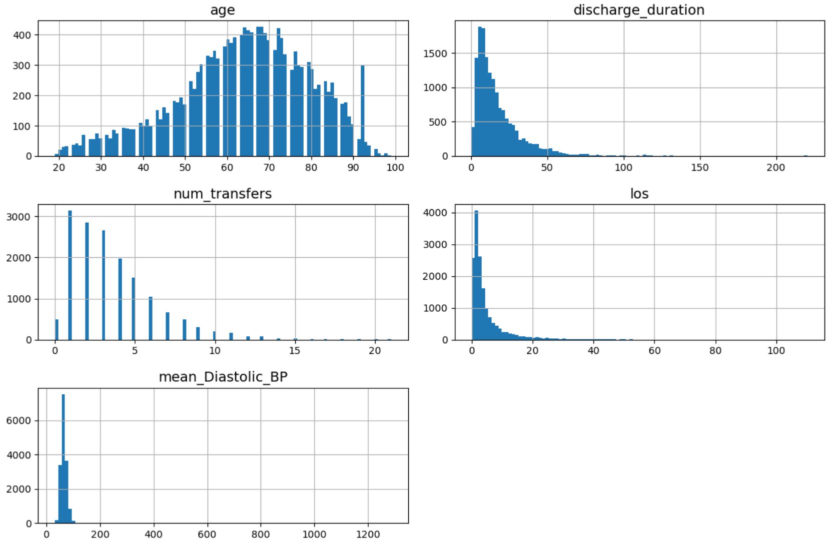


Figure 4: Among the five histograms, discharge duration, length of stay, and mean diastolic BP have heavy tails.

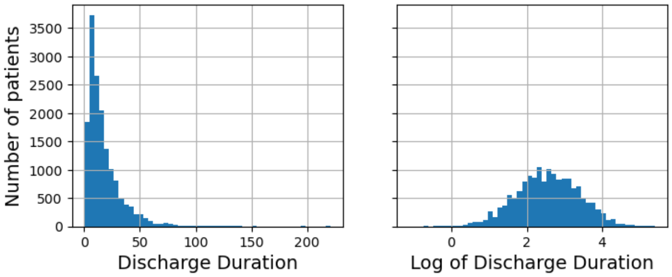


Figure 5: Example of discharge duration feature transformation with logarithm function.

Next, we investigated the linear correlation values between numerical features and is_readmission. Pearson correlation calculations show that discharge duration, number of transfers, and mean potassium have high values. Meanwhile, the scatter plot of those numerical features is depicted in Fig. 3.

In the next stage, the features will be subjected to a feature scaling process. However, histograms of some numerical features such as discharge duration, length of stay, and mean diastolic BP have a skewed right or heavy tail distribution (Fig. 4). Therefore, these features need to be transformed with a logarithmic function before they are standardized or normalized [10].

An example of a feature subjected to the logarithmic function, namely discharge duration, can be seen in Fig. 5. The total numerical features transformed by the logarithm function is 7 (seven), such as discharge duration, length of stay, mean diastolic BP, mean glucose, mean systolic BP, standard deviation of heart rate, and standard deviation of systolic BP.

After preprocessing the raw dataset, the number of samples becomes 19,967. Subsequently, the dataset is randomly divided into train and test sets with 80% and 20%, respectively. Training instances that are anomalies are discarded using the Isolation Forest algorithm [17, 10]. In the next step, the training instances are subjected to Z-scale feature normalization. Furthermore, various types of widely-used machine learning models in data science competitions, including linear, non-linear, ensemble, and boosting models, are employed and evaluated side by side. The linear models include Logistic Regression and Linear Discriminant Analysis (LDA), while the ensemble models encompass K-nearest neighbors (K-nn), Naïve Bayes, Decision Tree, and Support Vector Machines (SVM). Among the ensemble methods are the Bagging classifier with multiple Decision Trees, Random Forests, and Extra Trees. Lastly, the boosting models comprise Adaboost, Stochastic Gradient Boosting, and Light Gradient Boosting Machines (LightGBM).

As we employ numerous models, we opt to perform a two-stage evaluation process. Initially, all models are assessed using scikit-learn's default settings, and their performances are compared utilizing the stratified 10-fold cross-validation technique. Subsequently, we selected the top three models and conducted hyperparameter tuning on them.

The selected metric for evaluation is balanced accuracy, which offers an advantage by mitigating inflated accuracy due to imbalanced datasets. Moreover, the balanced formula is given as follows:

$$\text{balanced\_accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \qquad (2)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Equation (2) can alternatively be interpreted as the average recall across all classes, meaning that in a balanced dataset, the balanced accuracy aligns with the accuracy value.

2.2.1 Global Model-Agnostic Methods

Once we assess the performance of machine learning models, we proceed to examine the features that that substantially influence the prediction outcomes. Specifically, we employ two different techniques that are independent of the specific model being used to evaluate the significance of various features. These techniques include Permutation Feature Importance (PFI), which measures how much a model's prediction changes when a feature is permuted and SHAP values, which attribute the prediction to each feature.

Firstly, we employ Permutation Feature Importance (PFI) [9, 21] to gauge the importance of the features in Tabs. 1 and 2. Basically, PFI assesses how much the prediction error of the model rises when the values of a feature undergo permutation, effectively disrupting the correlation between the feature and the actual target value. A feature is significant if randomly changing its value leads to a larger error in the model's predictions. This condition indicates that the model uses and depends on that feature when making its predictions. On the other hand, a feature is unimportant if randomly altering its values does not impact the model's prediction error. This condition suggests that the model does not rely on that particular feature when generating predictions.

Secondly, SHAP (SHapley Additive exPlanations) is a technique employed to explain the prediction made by a model for a specific data instance [19]. SHAP calculations are based on an optimal game theory concept called Shapley values. These numerical numbers explain how each feature and its corresponding value influence or contribute to the prediction made by the model. Specifically, we are interested in understanding how much each individual feature value contributes to the prediction relative to the average or expected prediction value. Assume that $h(x)$ represents the prediction made by a model for a particular instance x,

$$h(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{3}$$

where $x_1, x_2, \cdots, x_n$ are feature values and our goal is to compute the contributions of each component or feature that led to that prediction. The contribution, represented by ϕ_j, of the j-th feature to the prediction $h(x)$ [21] is defined as follows:

$$\phi_j(h) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j)$$

where $E(\beta_j X_j)$ represents the average effect estimate for the j-th feature. If we add up all the contributions from the feature values of a single instance, the result is as follows:

$$\begin{aligned} \sum_{j=1}^{n} \phi_j(h) = \sum_{i=1}^{n} \beta_j x_j - E(\beta_j X_j) &= (\beta_0 + \sum_{i=1}^{n} \beta_j x_j) - (\beta_0 + \sum_{j=1}^{n} E(\beta_j X_j)) \\ &= h(x) - E(h(X)) \end{aligned} \tag{4}$$

Equation (4) shows that the sum of the contributions from all feature values for a given instance is equal to the predicted value for that instance minus the average predicted value across all instances. The principle behind the SHAP (SHapley Additive exPlanations) method for determining feature importance is straightforward: features

Table 3: Results of stratified 10-fold cross-validation of all models. The models marked with the † symbol perform the highest.

	Balanced Accuracy
Linear Models	
Logistic Regression	50.81%
Linear Discriminant Analysis	50.85%
Non-linear Models	
K-nn	56.52%
Naïve Bayes	50.92%
Decision Tree†	**61.99%**
SVM	50.45%
Ensemble Models	
Bagging Classifier	56.49%
Random Forests†	**58.34%**
Extra Trees†	**58.90%**
Boosting Models	
Adaboost	57.65%
Stochastic Gradient Boosting	56.70%
LightGBM	55.54%

Table 4: Parameters of models tested in hyperparameter tuning

Model	**Hyperparameter**	**Value Range**
Decision Tree	- `criterion` - `splitter`	[gini, entropy, log_loss] [best, random]
Extra Trees	- `n_estimators` - `criterion`	[100, 300, 900, 1200] entropy
Random Forest	- `n_estimators` - `criterion`	[100, 300, 900, 1200] entropy

Table 5: Performance of the best three models on test set

Model	**Best Hyperparameter**	**Balanced Accuracy**
Decision Tree	`criterion` : log_loss - `splitter`: best	61.98%
Extra Trees	- `n_estimators` : 100 - `criterion`: entropy	61.85%
Random Forests	- `n_estimators`: 100 - `criterion`: entropy	**63.70%**

that have high Shapley values are considered to be influential and carry significant weight in the model. The feature importance values can be derived by taking the average of the absolute Shapley value contributions for each individual feature across all the data points,

$$I_j = \frac{1}{n}\sum_{i=1}^{n}\left|\phi_j^{(i)}\right|.$$

3 Results and Discussion

Tab. 3 displays the results of the balanced accuracy evaluation across all models in the stratified 10-fold cross-validation. Tree-based machine learning models display superiority compared to other models of different types. This contradicts a recent literature study [6] suggesting that Gradient Boosting Machines, in this case LightGBM, typically outperform other algorithms in terms of Receiver Operating Characteristic-Area Under Curve (ROC-AUC) performance, which is analogous to balanced accuracy.

Afterward, we select the top three models from Tab. 3 for hyperparameter tuning. The hyperparameter ranges experimented with on these three models are illustrated in Tab. 4.

Once the hyperparameter tuning process completes and identifies the optimal hyperparameter settings for each model, the three algorithms are trained using these best parameters on the entire training set and then evaluated on the test set. Tab. 5 indicates that Random Forests display higher accuracy compared to both Decision Trees and Extra Trees. Additionally, we present the ROC-AUC curves for Decision Tree, Extra Trees, and Random Forests in Figs. 6, 7, and 8, respectively. The analysis of these three figures reveals a rise in the True Positive Rate (Recall) according to the model sequence: Decision Tree, Extra Trees, and Random Forests respectively; conversely, there is a decline in the False Positive Rate (Fallout) across these models as well.

Additionally, we assess the significance of features in these three models. A feature importance is determined by the cumulative reduction in both the mean and standard deviation of impurity across each tree [5]. A substantial reduction in impurity when employing a feature signifies a high information gain associated with that feature. Consequently, such features play a significant role in determining the class or label of instances. Fig. 9 illustrates a comparison of the top ten important features among the three models. It is evident that the three most significant features in all models are consistent: average hematocrit count, average platelet count, and the duration between a patient's admission and discharge. However, there is a difference in the order of importance between platelet count and discharge duration for the Extra Trees model compared to the other two models.

Pedregosa et al. [23] pointed out that calculating feature importance based on impurity measures can potentially be misleading for a feature with many categorical values. Therefore, we additionally examine which features are truly most influential for predicting the class by utilizing Permutation Feature Importance (PFI) [9, 21]. PFI is a model-agnostic global interpretation method for machine learning models, meaning it can be applied to analyze any model regardless of the algorithm. The computed values of PFI for the different features are visualized in Fig. 10.

The Permutation Feature Importance (PFI) analysis indicates that the average hematocrit count, average sodium level, average platelet count, and the length of hospital discharge duration are identified as the most important features for the models. These top important features align with those highlighted in the feature importance plots shown in Fig. 9. Interestingly, the binary feature `race_OTHERS`, which indicates whether a patient belongs to other racial groups besides the major ones, is unexpectedly identified as an important feature for making predictions by the models.

Next, we analyze feature importance by computing the SHAP values. Specifically,

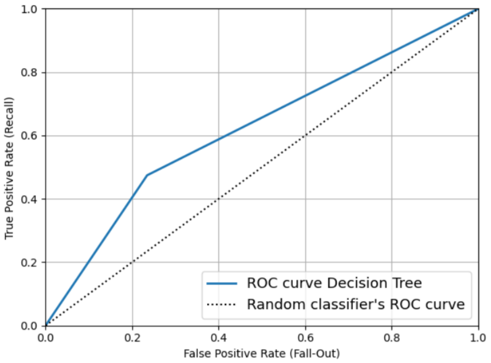


Figure 6: ROC-AUC of the Decision Tree model.

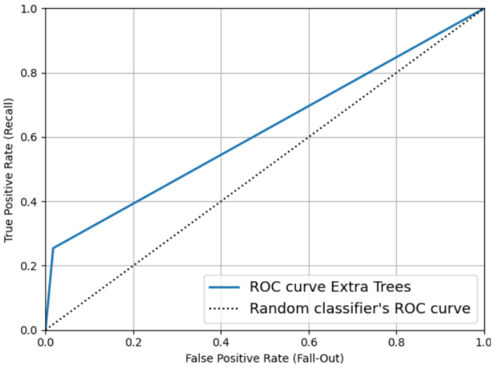


Figure 7: ROC-AUC of the Extra Trees model.

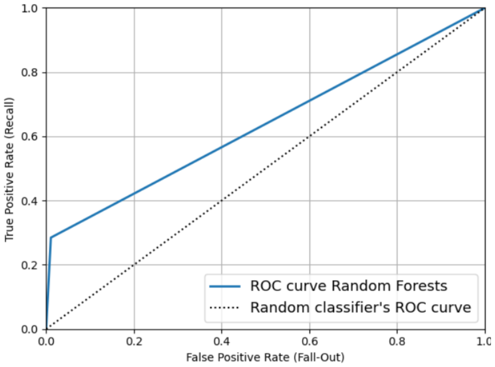
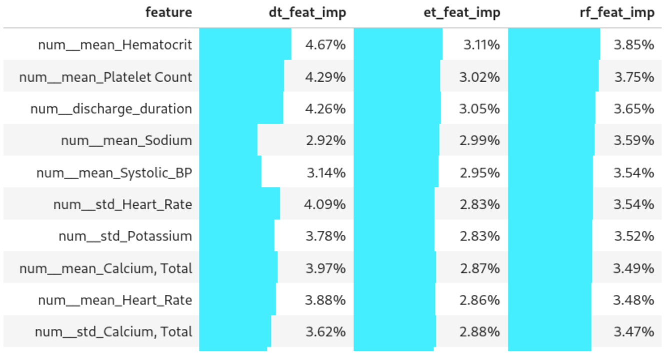
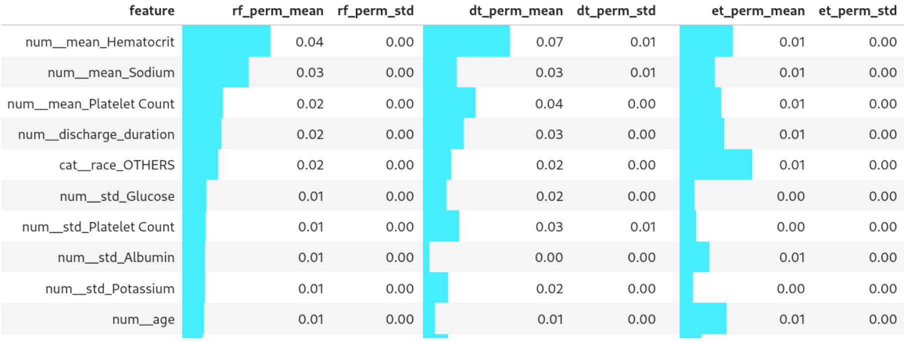


Figure 8: ROC-AUC of the Random Forests model.



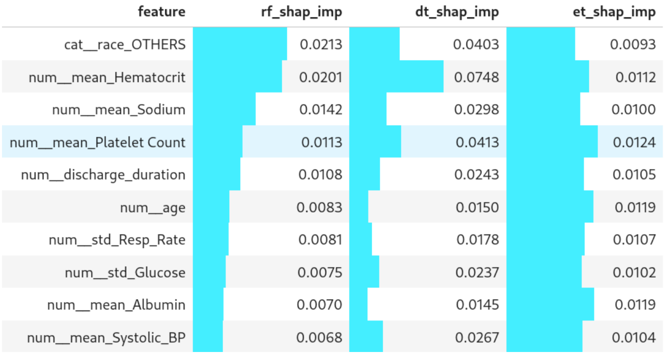
feature	dt_feat_imp	et_feat_imp	rf_feat_imp
num__mean_Hematocrit	4.67%	3.11%	3.85%
num__mean_Platelet Count	4.29%	3.02%	3.75%
num__discharge_duration	4.26%	3.05%	3.65%
num__mean_Sodium	2.92%	2.99%	3.59%
num__mean_Systolic_BP	3.14%	2.95%	3.54%
num__std_Heart_Rate	4.09%	2.83%	3.54%
num__std_Potassium	3.78%	2.83%	3.52%
num__mean_Calcium, Total	3.97%	2.87%	3.49%
num__mean_Heart_Rate	3.88%	2.86%	3.48%
num__std_Calcium, Total	3.62%	2.88%	3.47%

Figure 9: Feature importance across Decision Tree (`dt_feat_imp`), Extra Trees (`et_feat_imp`), and Random Forests (`rf_feat_imp`). The greater the percentage value of a feature, the greater the impact of the feature on determining the label of an instance.



feature	rf_perm_mean	rf_perm_std	dt_perm_mean	dt_perm_std	et_perm_mean	et_perm_std
num__mean_Hematocrit	0.04	0.00	0.07	0.01	0.01	0.00
num__mean_Sodium	0.03	0.00	0.03	0.01	0.01	0.00
num__mean_Platelet Count	0.02	0.00	0.04	0.00	0.01	0.00
num__discharge_duration	0.02	0.00	0.03	0.00	0.01	0.00
cat__race_OTHERS	0.02	0.00	0.02	0.00	0.01	0.00
num__std_Glucose	0.01	0.00	0.02	0.00	0.00	0.00
num__std_Platelet Count	0.01	0.00	0.03	0.01	0.00	0.00
num__std_Albumin	0.01	0.00	0.00	0.00	0.01	0.00
num__std_Potassium	0.01	0.00	0.02	0.00	0.00	0.00
num__age	0.01	0.00	0.01	0.00	0.01	0.00

Figure 10: The average of feature importance and standard deviation based on PFI across Random Forests (`rf_perm_mean` and `rf_perm_std` respectively), Decision Tree (`dt_perm_mean` and `dt_perm_std`), and Extra Trees (`et_perm_mean` and `et_perm_std`). A higher value for a particular feature corresponds to a larger error in the model's predictions. Consequently, higher-value features hold greater significance in determining the predicted output.



feature	rf_shap_imp	dt_shap_imp	et_shap_imp
cat__race_OTHERS	0.0213	0.0403	0.0093
num__mean_Hematocrit	0.0201	0.0748	0.0112
num__mean_Sodium	0.0142	0.0298	0.0100
num__mean_Platelet Count	0.0113	0.0413	0.0124
num__discharge_duration	0.0108	0.0243	0.0105
num__age	0.0083	0.0150	0.0119
num__std_Resp_Rate	0.0081	0.0178	0.0107
num__std_Glucose	0.0075	0.0237	0.0102
num__mean_Albumin	0.0070	0.0145	0.0119
num__mean_Systolic_BP	0.0068	0.0267	0.0104

Figure 11: The SHAP values across Random Forests (`rf_shap_imp`), Decision Tree (`dt_shap_imp`), and Extra Trees (`et_shap_imp`). A feature with a higher value contributes more significantly to the prediction of the output.

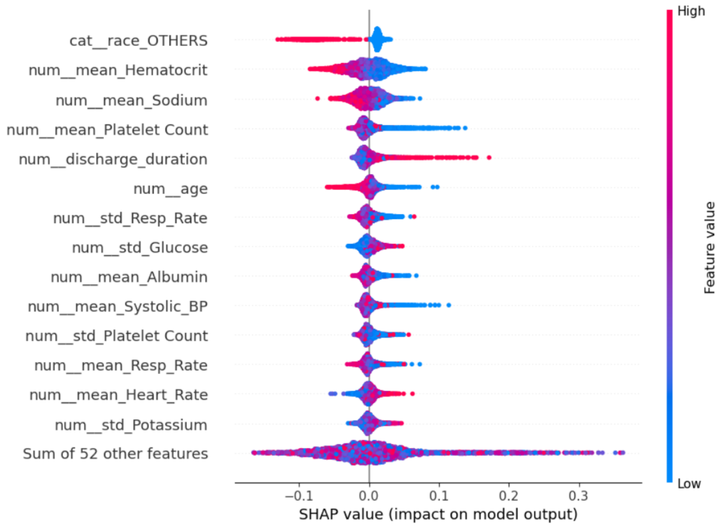


Figure 12: The SHAP summary plot for the Random Forest model indicates that values of average platelet count and hematocrit which are above zero decrease the probability of patient readmission, while a longer duration of hospital stay increases the likelihood of readmission.

we utilize KernelSHAP for Extra Trees and TreeSHAP for both Decision Tree and Random Forests. The SHAP feature importance plot in Fig. 11 corroborates that the average hematocrit count, sodium level, platelet count, and discharge duration are key contributing features for the model's predictions. Aligning with the Permutation Feature Importance (PFI) results in Fig. 10, the binary `race_OTHERS` feature is also found to influence the predictions. However, for the Extra Trees model, the SHAP values across features are relatively balanced, suggesting that multiple features contribute almost equally to the predictions.

The SHAP Beeswarm plot in Fig. 12 provides an overview of the SHAP values across all features. This plot is shown specifically for the Random Forest model, as it was determined to be the best-performing model in the experiments.

4 Conclusion

The study aims to predict patient readmission, framing it as a classification problem using the MIMIC IV dataset. Feature extraction from the dataset resulted in 27 numerical features and 7 categorical features. Several machine learning models were trained on the training set using stratified 10-fold cross-validation. The three best-performing models underwent hyperparameter tuning. The best model after hyperparameter tuning was the Random Forest model, which achieved a balanced accuracy of 63.7% when trained on the entire training set and tested on the test set.

Furthermore, this research delves deeper into identifying the key features that significantly influence the prediction model's outcomes. In addition to evaluating feature importance through the cumulative reduction in both the mean and standard deviation of impurity, the study employs two global model-agnostic techniques: permutation feature importance (PFI) and SHapley Additive exPlanations (SHAP). The findings from these three approaches were consistent, indicating that the average levels of hematocrit, sodium, and platelets in the blood, along with the duration between patient registration and hospital discharge, are crucial features that impact the prediction results.

Acknowledgement

The authors would like to thank Maranatha Christian University for the support and resources that made this research possible. This research is supported by the Internal Grant from Research and Community Service Institute (LPPM), Maranatha Christian University which is regulated in the Rector's Decree Number:
013/SK/ADD/UKM/III/2023.

References

[1] Anshik. *AI for Healthcare with Keras and Tensorflow 2.0: Design, Develop, and Deploy Machine Learning Models Using Healthcare Data.* Apress, New Delhi, India, 2021.

[2] Rasha Assaf and Rashid Jayousi. 30-day hospital readmission prediction using mimic data. In *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6, 2020.

[3] S.S. Aubakirov, P. Trigo, and D.Zh. Ahmed-zaki. Building a model to predict classifier accuracy. *Eurasian Journal of Mathematical and Computer Applications*, 5(3):4–14, 2017.

[4] R.A. Berenson, R.A. Paulus, and N.S. Kalman. Medicare's readmissions-reduction program: a positive alternative. *New England Journal of Medicine*, 366(15):1364–1366, 2012.

[5] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[6] T. Chen, S. Madanian, D. Airehrour, and M. Cherrington. Machine learning methods for hospital readmission prediction: systematic analysis of literature. *Journal of Reliable Intelligent Environments*, 8(1):49–66, 2022.

[7] Alex G. C. de Sá, Daniel Gould, Anna Fedyukova, Mitchell Nicholas, Lucy Dockrell, Calvin Fletcher, David Pilcher, Daniel Capurro, David B. Ascher, Khaled El-Khawas, and Douglas E. V. Pires. Explainable machine learning for icu readmission prediction, 2023.

[8] A.M. Epstein, A.K. Jha, and E.J. Orav. The relationship between hospital admission rates and rehospitalizations. *New England Journal of Medicine*, 365(24):2287–2295, 2011.

[9] A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[10] A. Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly Media, Inc., 2022.

[11] A. Gupta and G.C. Fonarow. The hospital readmissions reduction program: learning from failure of a healthcare policy. *European journal of heart failure*, 20(8):1169–1174, 2018.

[12] Y. Huang, A. Talwar, S. Chatterjee, and R.R. Aparasu. Application of machine learning in predicting hospital readmissions: a scoping review of the literature. *BMC Medical Research Methodology*, 21(1):96, May 2021.

[13] I.I. James and V.I. Osubor. Hostile social media harassment: A machine learning framework for filtering anti-female jokes. *Nigerian Journal of Technology*, 41(2):311–317, 2022.

[14] A.E.W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T.J. Pollard, B. Moody, B. Gow, L-W H. Lehman, L.A. Celi, and R.G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023.

[15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T-Y Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[16] L. Li, L. Wang, L. Lu, and T. Zhu. Machine learning prediction of postoperative unplanned 30-day hospital readmission in older adult. *Frontiers in Molecular Biosciences*, 9, 2022.

[17] F.T. Liu, K.M. Ting, and Z-H Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[18] Y-T Lo, J. C-h Liao, M-H Chen, Ch-M Chang, and C-T Li. Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms. *BMC medical informatics and decision making*, 21:1–11, 2021.

[19] S.M. Lundberg and S-I Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[20] P. Michailidis, A. Dimitriadou, T. Papadimitriou, and P. Gogas. Forecasting hospital readmissions with machine learning. In *Healthcare*, volume 10, page 981. MDPI, 2022.

[21] C. Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[22] A.A. Okandeji, O.F. Odeyinka, A.A. Sogbesan, and N.O. Ogunye. A comparative analysis of haemoglobin variants using machine learning algorithms. *Nigerian Journal of Technology*, 41(4):789–796, 2022.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] L. Serrano. *Grokking Machine Learning*. Manning Publications Co., 2021.

[25] M.S. Stefan, P.S. Pekow, W. Nsa, A. Priya, L.E. Miller, D.W. Bratzler, M.B. Rothberg, R.J. Goldberg, K. Baus, and P.K. Lindenauer. Hospital performance measures and 30-day readmission rates. *Journal of general internal medicine*, 28:377–385, 2013.

[26] K. Teo, C.W. Yong, J.H. Chuah, Y.C. Hum, Y.K. Tee, K. Xia, and K.W. Lai. Current trends in readmission prediction: An overview of approaches. *Arabian Journal for Science and Engineering*, August 2021.

[27] S. Wang and X. Zhu. Predictive Modeling of Hospital Readmission: Challenges and Solutions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(5):2975–2995, September 2022. Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics.

H. Bunyamin,
Maranatha Christian University,
Jl. Prof. drg. Surya Sumantri, M.P.H. No. 65,
Bandung, West Java, Indonesia
Email: `hendra.bunyamin@it.maranatha.edu`,

T. L. Wargasetia,
Maranatha Christian University,
Jl. Prof. drg. Surya Sumantri, M.P.H. No. 65,
Bandung, West Java, Indonesia
Email: `teresa.lw@med.maranatha.edu`,

J. Kasih
Maranatha Christian University,
Jl. Prof. drg. Surya Sumantri, M.P.H. No. 65,
Bandung, West Java, Indonesia
Email: `julianti.kasih@it.maranatha.edu`.

Received 12.05.2024, Accepted 25.09.2024