Al-Nahrain Journal of Science
ANJS, Vol. 29 (1), March, 2026, pp. 128 – 149

Explainable Machine Learning Models for Mortality Risk Prediction of Crimean-Congo Hemorrhagic Fever in Iraq

Tiba Zaki Abdulhameed[1,*], Rabia Al Mamlook[2,3], Haider Ali Hantoosh[4], Hasnaa Imad Al-Shaikhli[1], Yasir Younis Majeed[5], Suhad A. Yousif[1], Tasnim Gharaibeh[6]

[1]Department of Computer Science, College of Sciences, Al-Nahrain University, Baghdad, Iraq.
[2]Department of Business Administration, Trine University, IN, USA.
[3]Advanced Research Center for Plant and Complementary Medicine, University of Zawia, Libya.
[4]Public Health Department, Thi-Qar Directorate of Health, Thi-Qar, Iraq.
[5]Epidemiology, Ministry of Health, Baghdad, Iraq.
[6] Department of Computer Science, Kalamazoo College, Kalamazoo, MI, USA.

Article's Information

Received: 10.09.2025
Accepted: 04.12.2025
Published: 15.03.2026

Keywords:
Data Analysis,
Prediction Accuracy,
Outbreak,
Feature Importance Analysis,
Explainable AI.

Abstract

In mid-2022, Iraq experienced a massive outbreak of Crimean-Congo hemorrhagic fever (CCHF), resulting in high mortality rates. The outbreak began in Thi-Qar province and subsequently spread to other provinces. This research analyzes data collected from Thi-Qar province to investigate the key factors influencing patient life risk. This is accomplished by collecting a real dataset (HemoIraq24) and conducting a statistical analysis, followed by developing explainable patient outcome prediction models using several machine learning algorithms. The most important factors contributing to the decision of the predicted outcome are obtained using feature importance and SHAP techniques. In addition, a web-based application has been developed based on the best ML prediction model to assist healthcare providers in clinical decision-making. The ML algorithms tested include Decision Trees, Random Forests, Logistic Regression, Gradient Boosting, and K-nearest neighbor. The highest baseline prediction model accuracy achieved is 89%. Feature importance analysis and SHAP are utilized for further feature engineering, causing an enhancement of 3% in prediction accuracy, with up to 8% enhancement in F1 score. It is found that the main factor contributing to the patient outcome is the days in the hospital, which means that the healthcare given in the hospitals is strong enough and can handle the endemic. The dataset can help with future research and is available at: HemoIraq24 Dataset.

http://doi.org/10.22401/ANJS.29.1.13
*Corresponding author: tiba.zaki@nahrainuniv.edu.iq

1. Introduction

Crimean-Congo hemorrhagic fever (CCHF) is a viral disease caused by infection with a tick-borne virus and transmitted to humans either by infected tick bite or by direct contact with blood or tissue from infected humans and livestock [1][2]. The mortality rate caused by CCHF is between 10% and 40%[3]. CCHF poses a serious risk to public health that starts with non-specific symptoms like fever, muscle pain and headache. Those who survive may experience severe hemorrhaging from the mucous membranes, resembling petechiae under the skin, or internally, eventually resulting in shock, multi-organ failure and, in some cases death [4]. The first case was detected in Iraq in 1979 [5][6], and since Iraq has been considered endemic to CCHF (It consistently presents in a specific region or population). The cases were reported annually, but in mid-2022, Iraq experienced a huge outbreak that started in Thi-Qar province and then spread to other provinces, mainly in the middle and southern parts of Iraq, showing a surge in the number of cases. Meanwhile, Thi-Qar

reported the highest incidence of disease and a high mortality rate [7]. The CCHF dataset used in this research was obtained from the Zoonotic unit Thi-Qar Directorate of Health, from 2020 to 2024, in Iraq. It describes the epidemiological characteristics of CCHF cases that peaked during the first half of 2022. Each case is recorded in terms of age, gender, symptoms, residence, history of contact with another case, and history of contact with animals. In general, Machine Learning (ML) can significantly aid the healthcare sector in developing risk prediction models for various. [8]. Artificial Intelligence (AI) is increasingly used to improve the healthcare industry, aiming to improve patient recovery outcomes (mortality or recovery). However, many AI models produce good results, but they are working as a black box. The lack of explanation of the reasons behind the produced results may reduce its accountability and trust. Explainable AI (XAI) tries to reveal the causes and give explanations of the AI models' results. This helps humans to understand why specific decisions are suggested by AI models [9]. Recent research has produced an ML model for similar outbreaks in other countries to serve the field of health care. The CCHF dynamics in Turkey, a neighboring country to Iraq, are investigated by building a prospective prediction tool through a structured Gaussian process algorithm that reached an accuracy of 98.6% in risk assessment of specific geographic regions [10]. Research done on simulated data in China to predict hemorrhagic fever disease by policy-makers recommend considering the CNN-LSTM model as a time series analysis tool [11]. In addition, hemorrhagic fever with renal syndrome, which is widely endemic in China, reported 11,432 cases from 2010 to 2018 in Shandong. A boosted regression tree model for risk assessment was built with 91% accuracy [5]. To produce a ML risk prediction model for various diseases in many countries, an assembled model was trained by 43 diseases across 206 countries [12]. It is based on an assembly technique where the votes of five models (Neural Network, XGBoost, Logistic Boost, Random Forest, and Kernel SVM) are combined to achieve an accuracy of around 80% to 90% in risk prediction.

Other outbreak behavior, such as the Ebola Outbreak, was also studied using ML prognostic models [13]. Their Multivariate logistic regression model is conducted on 470 patients to predict the survival outcomes. The limited number of training cases produced a model with an accuracy range 64% to 74%. Other research on Ebola used a simulated outbreak dataset of 12,049 cases to mimic the Ebola epidemic that hit West Africa during 2013-2016. The AUC curve did not exceed the 70th due to the dataset limitation [14]. A mobile application for clinical diagnosis and prediction based on machine learning was developed for the Ebola outbreak during 2014-2015. A total of 106 patients in Sierra Leone clinical cases were used to train the model. The dataset is characterized as incomplete, unbalanced, and not in conformity to standard. The WHO Ebola Interim Assessment Panel advised that enhanced data collection, reporting, and monitoring tools are crucial for improving responses in future outbreaks [15]. Table 1 summarizes the reviewed studies with advances of AI and machine learning dedicated to healthcare with stand still gaps. One of these gaps is models that are built for a specific region (like Turkey and China) cannot be generalized. Other models are trained on simulated or incomplete datasets that fail to reflect the real-world complex situations. Further, small training datasets lead the trained model to have lower predictive accuracy, as proved in logistic regression models. In addition, the model transparency lack and reduces trust, thus highlighting the need for Explainable Artificial Intelligence (XAI) solutions. Generally, pointing out these gaps through accurate data collection, validation, and explainable models would raise the AI credibility and applicability in the healthcare sector.

Although since 2020 the number of CCHF cases in Iraq has been increasingly recorded, to the best of our knowledge, no prior studies have developed models to predict the mortality risk in CCHF patients or to systematically analyze the key factors influencing disease severity. This study advances AI applications in epidemiology by combining predictive power with explainability, ensuring that the results are both accurate and interpretable. This article proposes an ML model to predict the mortality risk of CCHF in Iraq using data collected from 2020 to 2024 in the province of Thi-Qar. The main contributions achieved in research are:

- Publishing a real data set for Crimean-Congo Hemorrhagic Fever (CCHF) for Iraqi patients to help/support further/future research and development.
- Developing an Explainable AI (XAI)-based model to predict the mortality risk. This is provided as a web application for the healthcare provider.

Building XAI models that explain the feature(s) importance to identify key risk factors associated with Crimean-Congo Hemorrhagic Fever (CCHF) in Iraq.

Table 1. Summary of studies on healthcare AI and ML models

Author(s)	Objective	Methods	Findings	Limitations
Sharma Mukta (2023)	XAI for healthcare trust	XAI techniques	Improves trust in AI models	XAI is not generalizable to all models
Ak C. (2020)	CCHF risk in Turkey	Gaussian Process	98.6% accuracy in regional risk	Region-specific only
Wang Z. D. (2024)	Forecast hemorrhagic fever	CNN-LSTM	Useful for policy-maker prediction	Based on simulated data
She Kaili (2021)	Hemorrhagic fever risk in China	Boosted regression trees	91% accuracy in risk assessment	Region/disease-specific
Zhang Tianyu (2024)	Multi-disease risk prediction	Ensemble (NN, XGBoost, RF, SVM)	80% to 90% accuracy across 43 diseases	Lacks specificity per disease
Colubri A. (2019)	Ebola survival prediction	Logistic regression	64% to 74% accuracy (limited data)	The small dataset affected the results
Forna Alpha (2021)	Ebola dynamics (simulated)	ML on simulated data	AUC below 70	Low-complexity simulated data
Colubri A. (2016)	Mobile ML for Ebola	App w/ 106 clinical cases	Emphasized better monitoring	Incomplete, unbalanced data

Section 2 provides a background on the applied ML algorithms is explained. The methodology is included in section 3. The results are presented in Section 4. In section 5, the web-based application and the ML prediction model used are explained. Discussion is proposed in section 6. Finally, the conclusions are given in Section 7.

2. ML Algorithm Background

This section provides an overview of the ML algorithms that are applied in this research.

2.1. Random Forest Algorithm

Bagging (bootstrap aggregation) is a statistical method of reducing the variance of a prediction function by creating multiple datasets via resampling and then aggregating the predictions. This technique works best with high-variance, low-bias models such as decision trees built for classification problems. Random Forest is bagging with a twist: it builds an ensemble of decorrelated trees and averages their predictions. Random Forest's key innovation is the reduction of correlation between trees, which enhances predictive power while avoiding variance increases. This method introduces randomization in terms of selecting features and data subsets used for tree construction. Random Forest often achieves performance comparable to boosting methods, while being easier to train and tune in many cases. As a result of its robustness and versatility, Random Forest has now been implemented in many software packages, and it is one of the most used ML algorithms across different disciplines [16].

2.2. Gradient Boosting Algorithm

Boosting is a family of machine learning algorithms that combines multiple weak learners create a single strong learner by improving their performance. Where a weak learner has a performance slightly better than a random guess, a strong learner has a performance of nearly optimal predictive accuracy [17]. Gradient Boosting is a supervised machine learning algorithm that aims to minimize the error of any differentiable loss function by iteratively reducing it. Using a single regression tree as the base learner, the algorithm builds a new regression tree to predict the residuals in each iteration step, which allows for a focus on improving predictions by correctly predicting the residuals. The approach starts with some baseline prediction and incrementally adds corrections to minimize some known loss functions, like the mean squared error. Single regression trees are well-liked due to their interpretable and computationally efficient predictions, as they can segment the data space into separate regions. While a single regression tree can be too simplistic and restricts the potential of the model to learn complex behavior in the data, it serves as a good stepping block for boosting more sophisticated techniques. Gradient Boosting applies to most loss functions and provides a natural framework for robust regression, so it is a widely used regression tool [18].

2.3. Decision Tree-Based Algorithm

A decision tree is a supervised machine learning algorithm used for both classification and regression tasks. It is robust, interpretable, and works by recursively partitioning the input space according to attribute selection. C4 is the implementation of J48 C4.5, one of the most common decision tree classifiers, which allows efficient processing of both categorical and numerical data, implements pruning techniques reducing overfitting problems, and does not make any assumptions regarding the missing values [16]. The J48 algorithm is considered one of the baselines in decision tree learning, however, most decision trees now build hierarchical models that, through splits that minimizes impurity at each node, provide an intuitive model structure for classification and regression tasks [19]. Their ability to overfit complex datasets has led to the advent of ensemble methods [15], like Random Forest and gradient-boosted trees, which enhance predictive accuracy by bagging and boosting multiple decision trees. There have been theoretical advancements in decision trees, making them more efficient, making recent optimizations unique, and continuing to reduce complexity while ensuring generalization. Hence, decision trees are still at the forefront of machine learning research and applications, maturing over the years with the evolution of data science techniques [20].

2.4. Logistic Regression

Logistic regression, sometimes called the logit model, is a statistical model that, in its basic form, uses a logistic function to model the relationship between a binary (dichotomous) dependent variable and one or more continuous independent variables. It uses a logistic function to fit input data (cases and controls in epidemiological studies) and models the probability of an outcome occurring. The relationship is quantified by regression coefficients that show the directness of this relationship, as well as how much each predictor variable contributes to the probability of the output. Logistic regression is divided into two types. If the dependent variable is dichotomous, then a binary logistic regression is applied; otherwise, if the dependent variable has more than two different categories, a multinomial logistic regression is applied. Logistic regression is widely used in medical research, social science, and machine learning for predicting outcomes and assessing risk [21].

2.5. K-Nearest Neighbors

The k-nearest neighbors (k-NN) classifier is a traditional non-parametric method used for classification tasks. It determines the class of an unknown instance (a point in the feature space) by calculating its distance from instances in the dataset of the classifier. Usually, Euclidean distance is used as a metric to measure similarity. After computing distances, the instance is assigned the most common class among its k nearest neighbors (k being a pre-defined integer). This makes k-NN flexible and powerful, thus well-suited for pattern recognition and machine learning tasks [22].

2.6. Evaluation Metrics

The following prediction metrics are used to evaluate the prediction performance:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \dots (1)$$

$$Precision = \frac{TP}{FP + TP} \quad \dots (2)$$

$$Recall = \frac{TP}{FN + TP} \quad \dots (3)$$

$$F1score = \frac{2TP}{2TP + FP + FN} \quad \dots (4)$$

Where TP, TN, FN, and FP stand for the total number of True Positive, True Negative, False Negative, and False Positive, respectively. A positive in this research refers to a mortal case, and a negative refers to a recovered case. A TP is the number of mortal cases that are correctly predicted, and a TN is the number of recovered cases that are also correctly predicted. However, FP is the number of recovered cases that are incorrectly classified as mortal cases, and FN is the number of mortal cases that are incorrectly predicted. The Confusion Matrix (CM) is also generated to visualize the performance, where a heat map, based on the TP, TN, FP, and FN, is generated. It illustrates a visualization of the prediction model performance. Higher values of TP and TN indicate better performance.

3. Methodology

The methodology used in this article provides an advancement of an accurate, interpretable, and optimized model for mortality risk assessment. The process begins with data collection and the selection of significant features, followed by Pre-processing and splitting into training and testing sets. Baseline models, including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), and K-Nearest Neighbors (KNN), are trained, and their accuracy evaluated. Feature importance analysis is produced for the best-performing model. The correlation matrix that shows linear relationships is also generated. Following, feature engineering is applied to enhance the dataset, and

the top-performing models, RF, LR, and GB, are optimized via the engineered data. Accuracy is re-evaluated, and the final model goes through a detailed feature importance analysis and statistical analysis to identify the important predictors. The dataset used is explained in subsection 3.1, while a detailed description of the conducted analysis method with the process workflow graph is presented in subsection 3.2.

3.1. Thi-Qar Hemorrhagic Fever Data Description:

The dataset was collected at the medical institutions of Thi-Qar Directorate of Health/Public Health Department/Iraqi Ministry of Health since 2018. Only one instance was recorded during 2018. The number of instances started to increase since 2021. The total instances in the given dataset are 360. It consists of 39 fields of information as listed

1 YEAR
2 Month
3 Month number
4 date of investigation
5 date of admission
6 hospital name
7 age
8 age groups
9 gender
10 occupation
11 source of infection
12 province
13 district
14 type of area
15 Are there animals near the residence?
16 Are there ticks on animals or in barns?
17 village
18 Was an animal slaughtered during the current month inside the house?
19 Is there any contact with a similar case?
20 Are there any confirmed previous infections in the same residential area?
21 clinical features
22 date of onset
23 fever
24 gum bleeding
25 conjunctivitis
26 subcutaneous bleeding
27 bleeding from injection sites
28 bleeding from body orifices
29 bloody diarrhea
30 hematuria
31 date of blood sample taken
32 days in hospitalization
33 date of discharge
34 date of death
35 date of blood sample receipt
36 blood sample status
37 lab result
38 duration between sample taken and lab received
39 patient recovery outcome

The dataset statistics show that about 80% of the patients' outcomes shows recovery, while 20% are mortal (as death in Figure 1). The male cases are 25% higher than the female (Figure 2a). The most frequent ages are between 20 to 55 (Figure 2b). In addition, most patients are from rural areas, which is about 25% more than cases from urban areas, and 55% more from suburban areas (Figure 2c). The source of infections is mainly through contact with animals and raw meat (Figure 2d).

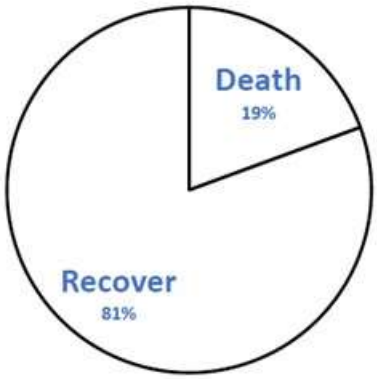

Figure 1: Classes of patients' outcomes.

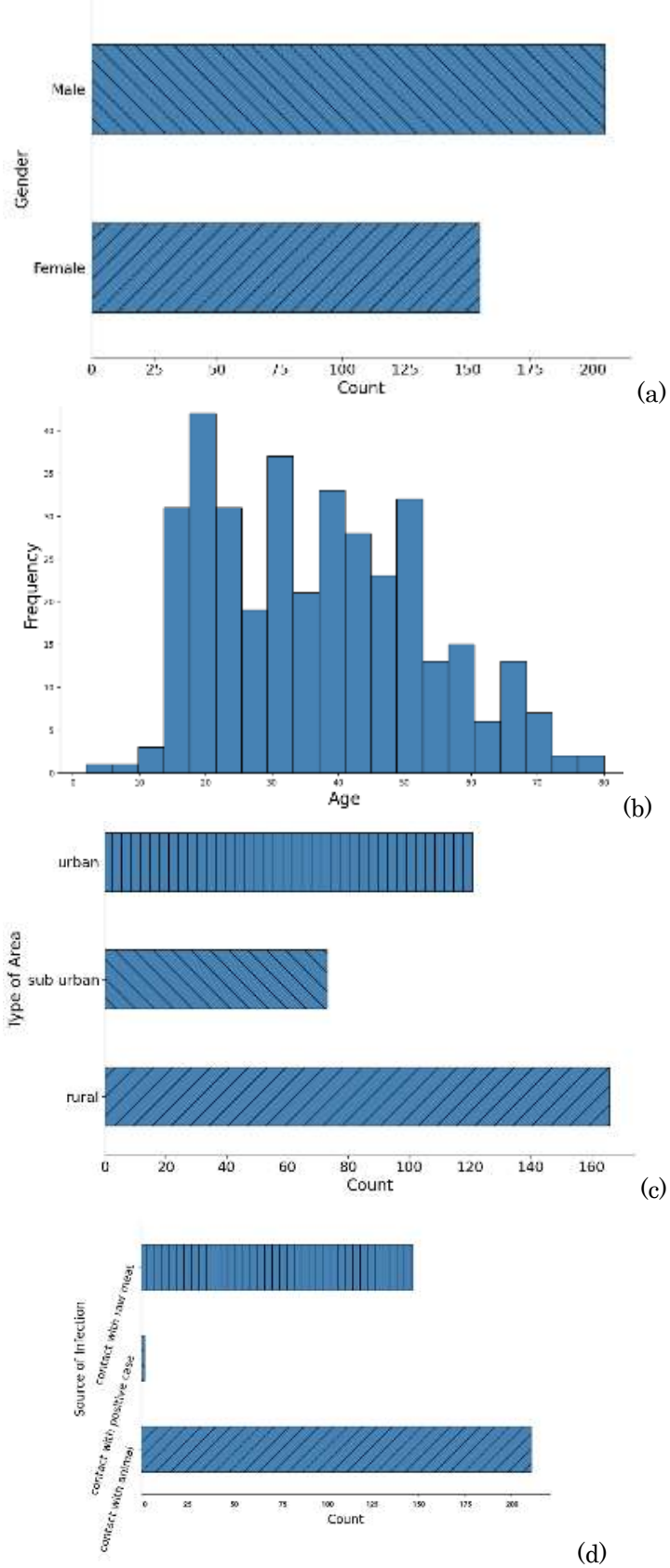

Figure 2: Summary of patient dataset distributions where (a) Patients' gender distributions, (b) Patients' age distribution, (c) Type of area distribution and (d) Source of infection distribution.

3.2. Proposed Systematic Risk Factors Assessment

The structured approach for developing machine learning models for Crimean-Congo Hemorrhagic Fever is shown in Figure 3. It begins with data collection and pre-processing, followed by training five baseline models (e.g., RF, DT, LR, and k-NN) to evaluate initial performance. Feature importance analysis will then be generated for the best model to identify the key variable fields. Feature engineering is then applied to enhance the dataset, and optimized models are developed. The top-performing model is re-analyzed for feature importance, to ensure accuracy and interpretability. This process is ideal for applications like disease prediction and decision support systems.

1. Feature Selection:

The health care expert chooses a list of fields that need to be studied, as listed in Table 2. Several other columns, related to clinical metadata, are removed to reduce data complexity. In addition, the health care expert computed two new features, OnsetToAdmission and OnsetToOutcome. OnsetToAdmission is the date of admission to the hospital minus the date of onset. Moreover, OnsetToOutcome is calculated as the date of discharge (in the case of a recovered patient) minus the date of onset. On the other hand, if the patient outcome's is mortal, it is calculated as the date of mortal minus the date of onset. These two features show the history of illness duration.

2. Data Pre-processing:

The collected data needs some pre-processing to get it ready to be used as training and testing data for ML models. First, the categorical columns are mapped to classes. Then, the missing values are handled in two ways. The numerical and categorical (other than binary classes) are filled with the most frequent value of each column, such as ('age groups', 'age', 'OnsetToAdmission', 'OnsetToOutcome', 'days in hospitalization', and 'type of area', 'gender', 'occupation', 'source of infection', 'patient recovery While in the case of binary valued columns that represent specific symptoms ('bleeding from injection sites', 'bleeding from body orifices', 'gum bleeding', 'conjunctivitis', 'subcutaneous bleeding', 'bloody diarrhea', and 'hematuria'), empty cells are filled with zeros, which indicates that the patient was not having this symptom. Then, all categorical columns are converted to numerical format by applying label encoding. This is to ensure that the data is compatible with machine learning algorithms. The resulting data is ready to be used in effective training and testing of machine learning models. Finally, the dataset is split into 70% for training, while 30% of the instances are kept for the testing phase. Correlation matrix of the dataset is generated to reveal linear relationships of the feature to the patient recovery outcome.

Table 2. Collected features, types, and their descriptions. N for Numerical and C for Categorical data types.

Feature Name	Type	Description
Age Groups	N	Age group classification of the patient
Age	N	Age of the patient in years
OnsetToAdmission	N	Days from symptoms onset to hospital admission
OnsetToOutcome	N	Days from symptoms onset to outcome
Days in Hospitalization	N	Total days spent in the hospital
Type of Area	C	Geographical area type (Urban, suburban, and rural)
Gender	C	Male or female
Occupation	C	patient's work
Source of Infection	C	animals, raw meat, or contact with positive case
Patient Recovery Outcome	C	Every recovery is 1, while mortal is 0)
Bleeding from Injection Sites	C	Binary value: 1 if symptom present, 0 if absent
Bleeding from Body Orifices	C	Binary value: 1 if symptom present, 0 if absent
Gum Bleeding	C	Binary value: 1 if symptom present, 0 if absent
Conjunctivitis	C	Binary value: 1 if symptom present, 0 if absent
Subcutaneous Bleeding	C	Binary value: 1 if symptom present, 0 if absent
Bloody Diarrhea	C	Binary value: 1 if symptom present, 0 if absent
Hematuria (bloody urine)	C	Binary value: 1 if symptom present, 0 if absent

outcome', and 'type of area').

3. Model Building (Baseline) and Evaluation:
Five ML models are applied. The reason behind choosing them is as follows

i. Decision Tree (DT) is applied because it is the simplest non-linear model,
ii. Logistic Regression (LR) is a linear classifier; its coefficients directly indicate the magnitude and direction of a feature's influence on the outcome
iii. Random Forest (RF), and Gradient Boosting (GB) models are ensemble techniques known to be state-of-the-art due to their ability to capture complex, non-linear feature interactions.
iv. k-NN is applied to offer a different way to look at the data structure, helping to confirm that results aren't just working for one type of model architecture.

All of them are applied with k-fold cross-validation. These models are evaluated using accuracy, precision, recall, and F1 score. After evaluation, the analysis step follows.

4. Feature Importance Analysis:
The most accurate models are considered for further analysis to reveal and explain the most important risk factors affecting the decision of the ML model. This is done through Feature importance and SHAP.

5. Feature Engineering:
The model explanation step above and the features correlation heatmap help the healthcare experts to optimize the models, this is done by selecting the most relevant feature that contributes to the final decision of the patient's infection severity. A correlation heatmap is generated for the new list of features to justify the how are the new list of features are related to each other.

6. Model Building (Optimized):
Rebuild the three most accurate ML models but with the new list of features (important features). A confusion matrix is generated for the models with the highest performance.

7. Explainable AI Analysis:
For the most accurate models, results are explained using feature importance. This step reveals the most important factors contributing to patient recovery. This final analysis is conducted utilizing the results from XAI and from statistical tools (correlation matrix and feature distribution by risk level).

4. Results

This section illustrates the performance of the proposed models in their baseline version, and after analyzing the most important factors using the explainable AI to achieve optimized models. The most important factors, that are presented below through the analysis refer to the key clinical and epidemiological variables that had the strongest influence on patient recovery outcome prediction according to the explainable machine learning models.

4.1 Baseline Models Performance

To assess and compare the performance of multiple models, Precision, Recall, F1-score, and Accuracy were evaluated and presented in four separate bar charts, as shown in Figure 4. The bar charts provide a clear visualization of the models' strengths and weaknesses across the different performance metrics, while the corresponding Table 3 lists the actual values for reference. When evaluating models in a healthcare context, Recall is the most critical metric, as it reflects the model's ability to identify high-risk patients correctly. A low Recall value indicates an increased risk of misclassifying high-risk cases, potentially delaying necessary life-saving treatments. Although most models' accuracy are comparable, when prioritizing patient safety, the recall metric is the most important among others. A low value of recall means a higher probability of misclassifying the high-risk patient, which may cause delay in life-saving treatment. Figure 4 shows that the FI-Score for DT is 42% and for the k-NN is 55%. Thus, both are too low to be considered. On the other hand, RF, LR, and GB models produced the highest performance resulting an F1-score of 65%. This shows that, RF, LR, and GB achieved better balancing between precision and recall making them more reliable for identifying critical cases. Since healthcare is a high risk to human environment, the recall metric is very crucial. RF achieved the highest recall of 75% reflecting its strength in capturing high risk cases. LR and GB produced comparable results, both achieving a recall value of 69%, which makes them reliable models. Additionally, considering the overall accuracy, all three models, RF, LR, and GB, reached high accuracy level with 88% for RF and 89% for both LR and GB. These three models (RF, LR, and GB) are the most promising models. As we move forward, a closer examination of these models will focus on identifying the critical factors that drive their predictions. This deeper understanding of their decision-making process will not only enhance their reliability but also ensure their

practical value in real-world applications, particularly where accurate and timely predictions are vital. Thus, these models are considered for analyzing the most important factors that affected their final decision.

4.2. Feature Importance and SHAP for the Best Baseline Models

Feature Importance charts shown in Figures 5, 6, and 7 for the three highest performing models RF, LR, and GB respectively. These figures can offer useful insights into the factors that have a significant impact on the final model's predictions. The number of days in hospitalization presented as the most critical variable in both the RF and GB models. This finding highlights the significance of hospitalization duration in patient outcomes.

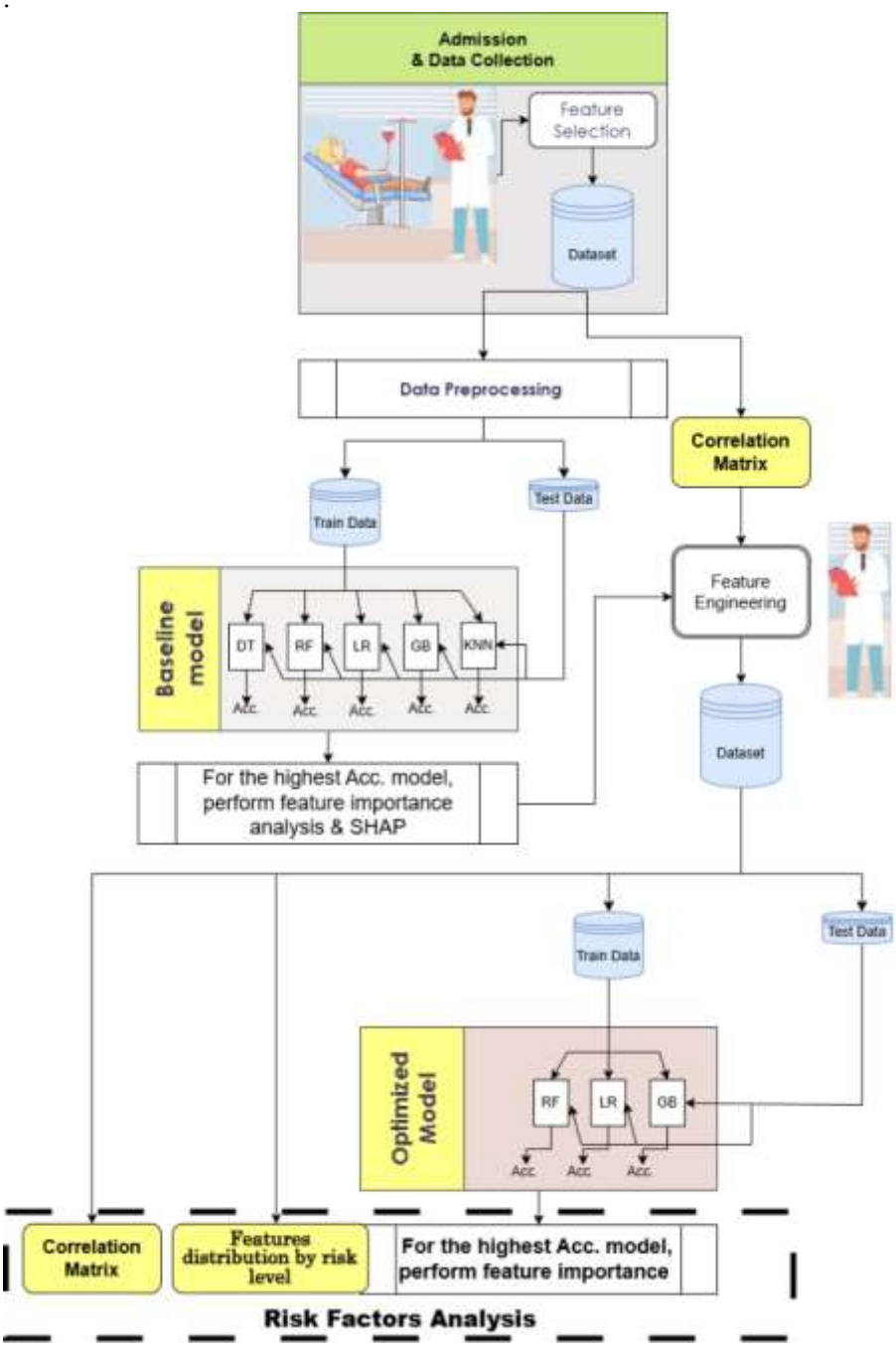

Figure 3: Process workflow

Table 3: Baseline performance metrics

Model	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
DT	33	56	42	77
RF	57	75	65	88
LR	61	69	65	89
GB	61	69	65	89
k-NN	53	56	55	86

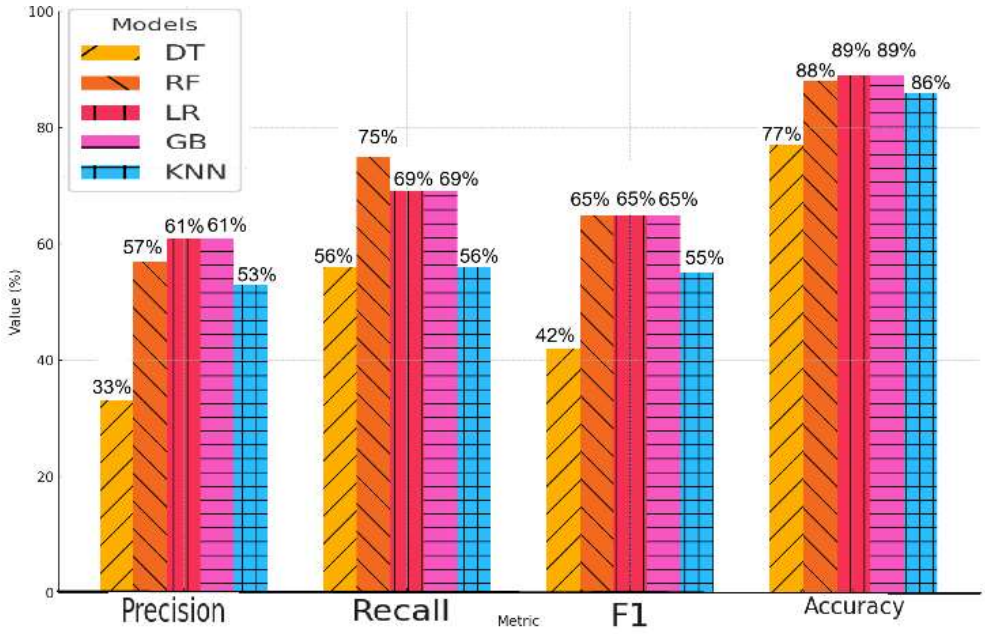

Figure 4: Performance of the baseline models in predicting the patient outcome.

Interestingly, in the LR model, days in hospitalization ranked as the third most important feature, highlighting a slight variation in how different algorithms weigh this factor. Looking at the Age factor, that is generally considered a significant demographic factor, it showed a disagreement of the levels of importance across the models. In the LR model, age was considered the least important feature, suggesting that it may not highly contribute to predicting outcomes. However, both the RF and GR models identified age as a highly important factor, indicating its relevance in understanding patient risk and health status. Moreover, conjunctivitis was consistently ranked as having minimal contribution to the final decision across all models. This agreement suggests that, despite being a potential symptom, conjunctivitis does not play a significant role in predicting patient recovery outcomes for this specific dataset. The importance of the remaining features varied across the models. This is due the differences of the models' inherent structures and learning mechanisms. To optimize model performance, healthcare experts will analyze these feature importance charts in conjunction with the feature correlation map. Moreover, considering Gradient Boosting model as a representative model, SHAP analysis is illustrated. The SHAP plots are shown in Figure 8, there are two realistic ways to understand SHAP plots. First, looking at the order of the features on the left side, the SHAP plots agreed with the feature importance produced in Figure 7.

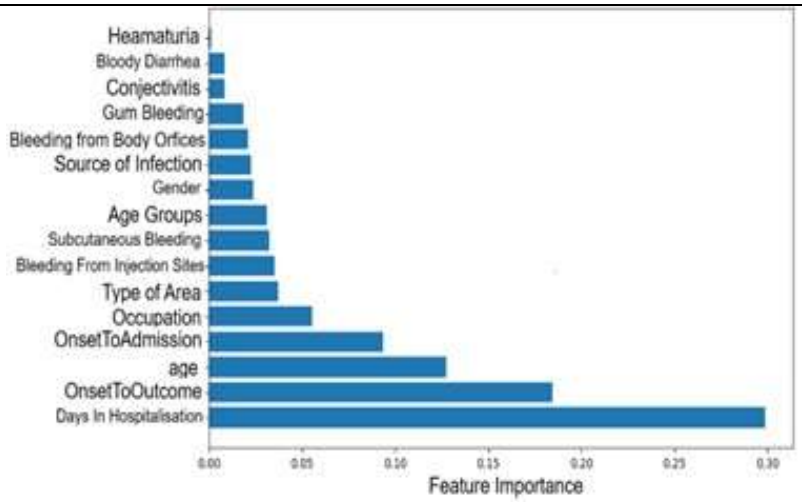

Figure 5: Random Forest feature importance.

Second, in each feature, the effect of its value to the mortal risk is shown, where each dot represents a patient from the test set. For each feature, the position of the dot on the horizontal axis indicates the SHAP value. A positive SHAP value indicates that the feature value increases the likelihood of predicting class 1 patient (recovery case), while a negative SHAP value decreases it.

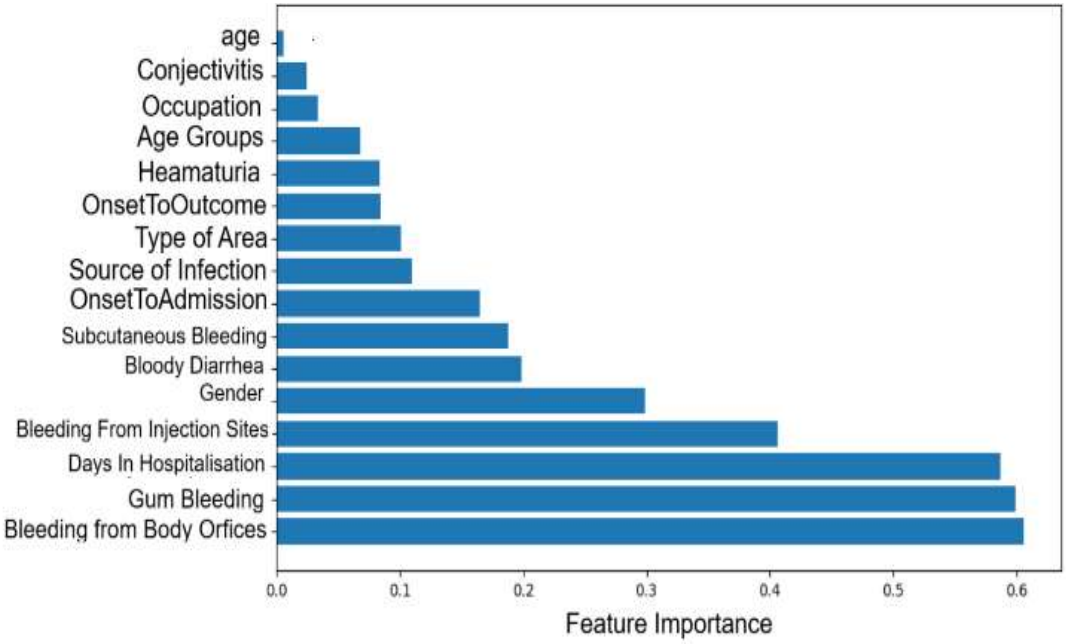

Figure 6: Logistic Regression feature importance.

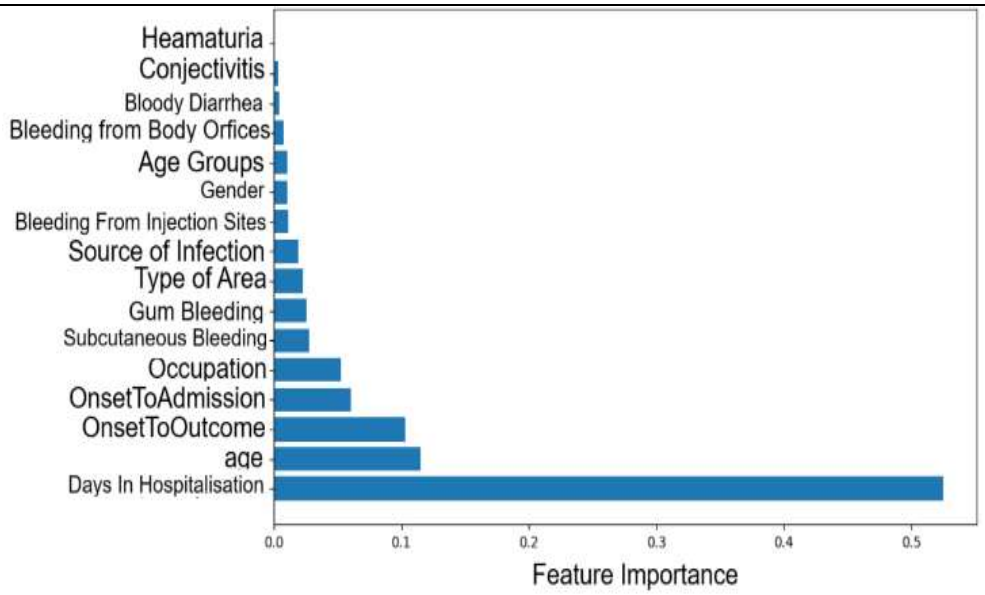

Figure 7: Gradient Boosting feature importance.

The red color is for high feature value, while the blue color indicates lower feature value. Thus the 'days in hospitalization' blue dots towards the left decreases the likelihood that patients would recover. For the age feature, smaller values (blue dots) on the right indicates higher likelihood to recover.

4.3. Feature Selection to Optimize the Models

The final maintained features are decided by the heath care experts after pointing out the most important features using Explainable AI tools (feature importance and SHAP), and with a deep analysis of the correlation matrix heatmap. The correlation matrix heatmap on Figure 9 highlights significant relationships between features in the dataset, providing insights into their influence on patient recovery outcomes. There is a strong positive correlation between age and age groups (0.86), confirming that both are closely aligned and one of them is enough for consideration. Likewise, the correlation between hematuria and bloody diarrhea is (0.55), indicating that these symptoms often co-occur, reflecting the severity of the disease.

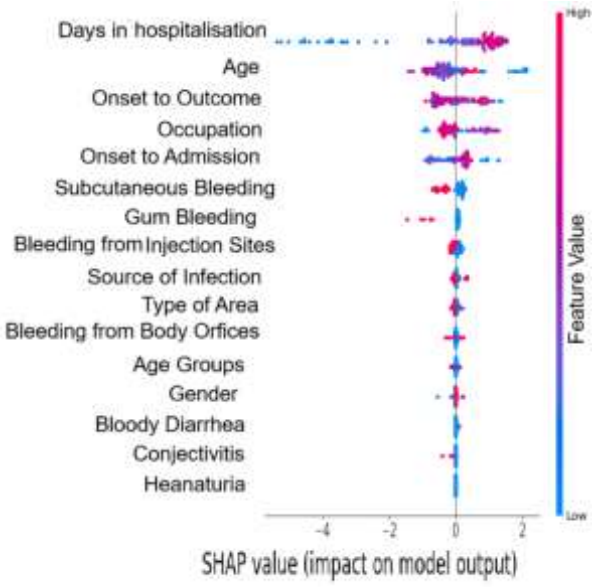

Figure 8: Gradient Boosting SHAP features analysis.

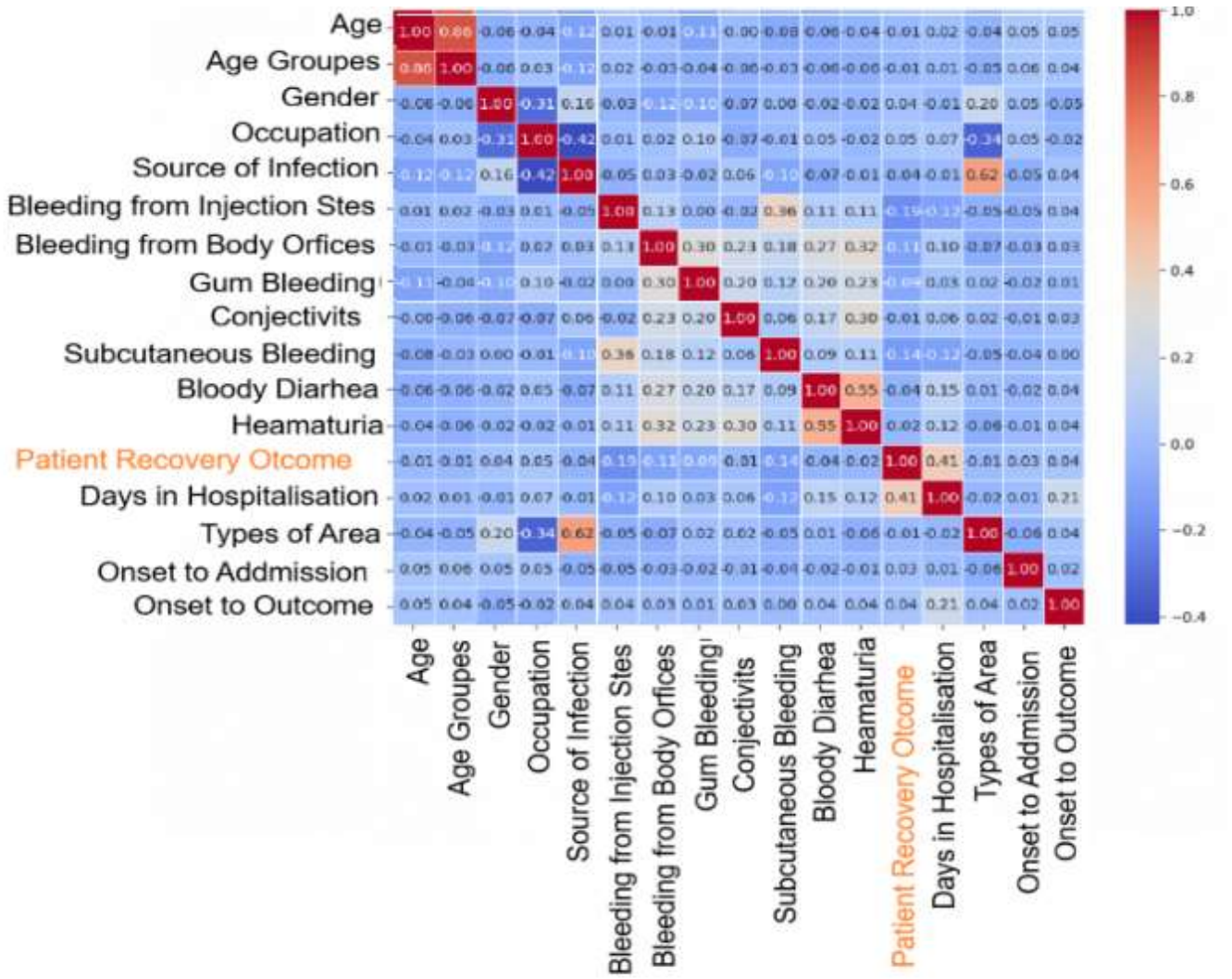

Figure 9: The correlation of the chosen features.

In contrast, negative correlations, such as between bleeding from body orifices and patient recovery outcomes (-0.19), reveal that patients experiencing such symptoms are more likely to experience worse outcomes, which may cause higher mortality risk. Another weak but significantly negative correlation is seen between age and patient recovery outcome (-0.02), suggesting that older patients may face slightly poorer prognosis. Key predictors like days in hospitalization show a positive correlation with patient recovery outcomes (0.41), indicating that extended hospital stays are associated with improved chances of recovery. The correlation between the type of area and source of infection (0.62) highlights the influence of geographical factors on disease transmission, likely due to variations in exposure to infected animals or environmental risks. Thus, according to the healthcare provider experts, a bleeding description column is created by summing all the binary columns indicating different bleeding symptoms (bleeding from injection sites + bleeding from body orifices + gum bleeding + conjunctivitis + subcutaneous bleeding). To simplify the analysis, the healthcare team decided to exclude various demographic factors, such as occupation, age group, type of area, source of infection, and gender, from consideration. This decision was made to sharpen the focus on the most relevant clinical indicators that impact patient recovery outcomes. Similarly by reducing the number of features analyzed, the team aims to enhance the clarity and precision of the model's predictions, the factors that are chosen to

conduct this analysis include age, bloody diarrhea, hematuria, days in hospitalization, onset to admission, onset to outcome, bleeding, and patient recovery outcome. This focused approach emphasizes the importance of prioritizing relevant clinical indicators to support effective decision-making in healthcare.

4.4. Optimized Model's Performance

After feature reduction, 8 selected features remains to be considered. The best models performances, listed in Table 4, are improved. The enhancement, shown in Figure 10, mainly affected the F1-score and accuracy. Optm.RF (Optimized Random Forest) is particularly suited to address the challenges of CCHF in Iraq due to its balanced and robust performance in key metrics. With its high precision and recall, this model ensures accurate identification of cases while minimizing false negatives and false positives, which are critical for the management of a life-threatening disease such as CCHF, the recall of the Optm.RF model is the highest among the other two models. This is especially valuable in public health context where missing a single case could result in further transmission and mortal outcomes. Furthermore, the high F1-score of the Optm.RF model provides an effective balance between precision and recall, ensuring accurate diagnoses.

Table 4. Final performance of the optimized models.

Model	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
Optm.RF	71	75	73	92
Optm.LR	73	69	71	92
Optm.GB	65	69	67	90

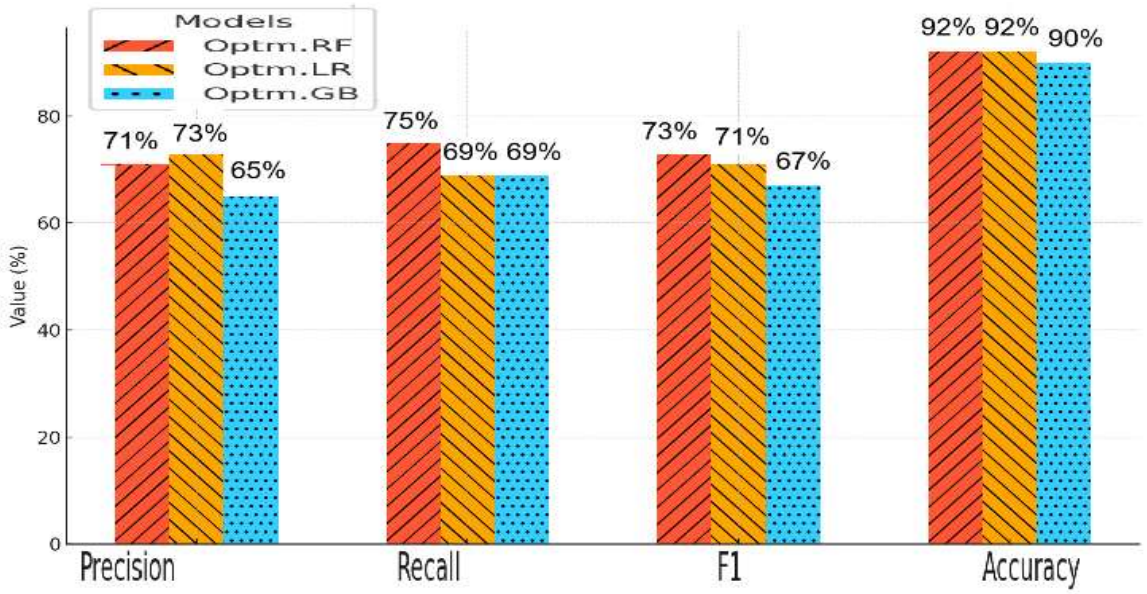

Figure 10: Patient's illness severity prediction performance of the optimized models.

The confusion matrix of the RF model (Figure 11) shows a good performance, with reduced classification errors. Five false negatives cases and four false positives make RF model reliable to detect critical conditions such as CCHF, hence the confusion matrix highlights its robustness, making it a practical choice for public health applications. Building an explainable AI prediction model helps to reveal the relations and the most important risk factors that contribute to patient recovery by showing how each feature contributed to the prediction of the patient recovery outcome. The ROC graph illustrated in Figure 12 shows the True Positive Rate (TPR or Sensitivity) against the False Positive Rate. The Area Under the Curve (AUC) reaches 0.89, 0.88, and 0,90 for the RF, LR, and GB models, respectively. In summary the overall ability of these models to distinguish between the high-risk and low-risk patients is revealed.

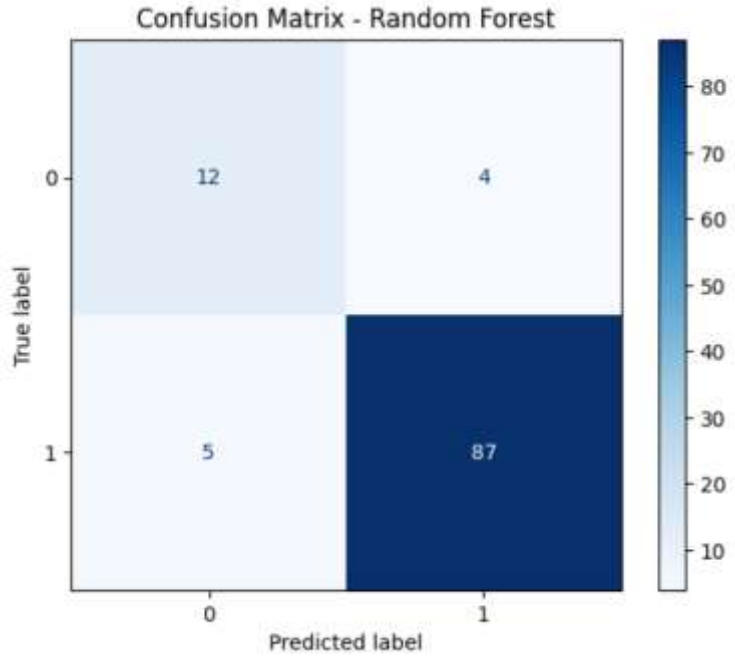

Figure 11: Confusion matrix for Optimized Random Forest results

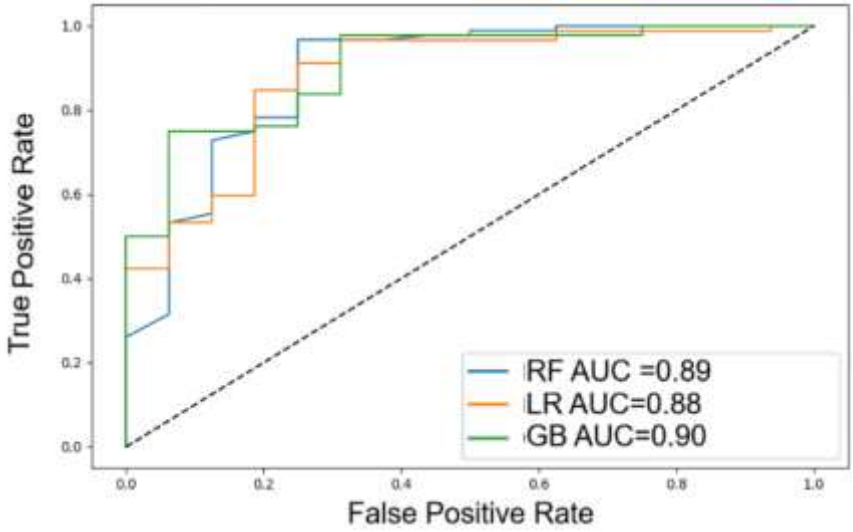

Figure 12: Area Under Curve (AUC) graph

The correlation matrix of the remaining columns is shown in Figure 13 The patient recovery outcome value is labeled 0 for mortal and 1 for recovery. The correlation between bleeding and patient recovery outcome is negative (-0.21); therefore, as the bleeding value increases, the patient recovery outcome becomes 0 (high risk of death). The same thing applies to age correlation to recovery outcome. The correlation value is (-0.02), which means elder patients are at a higher risk. Additionally, the correlation among all symptoms (bleeding, hematuria, and bloody diarrhea) is high (light orange colored). Moreover, bloody diarrhea and hematuria are highly related. The high positive correlation between days in hospitalization and patient recovery outcome indicates that longer duration of hospitalization increases the survival probability.

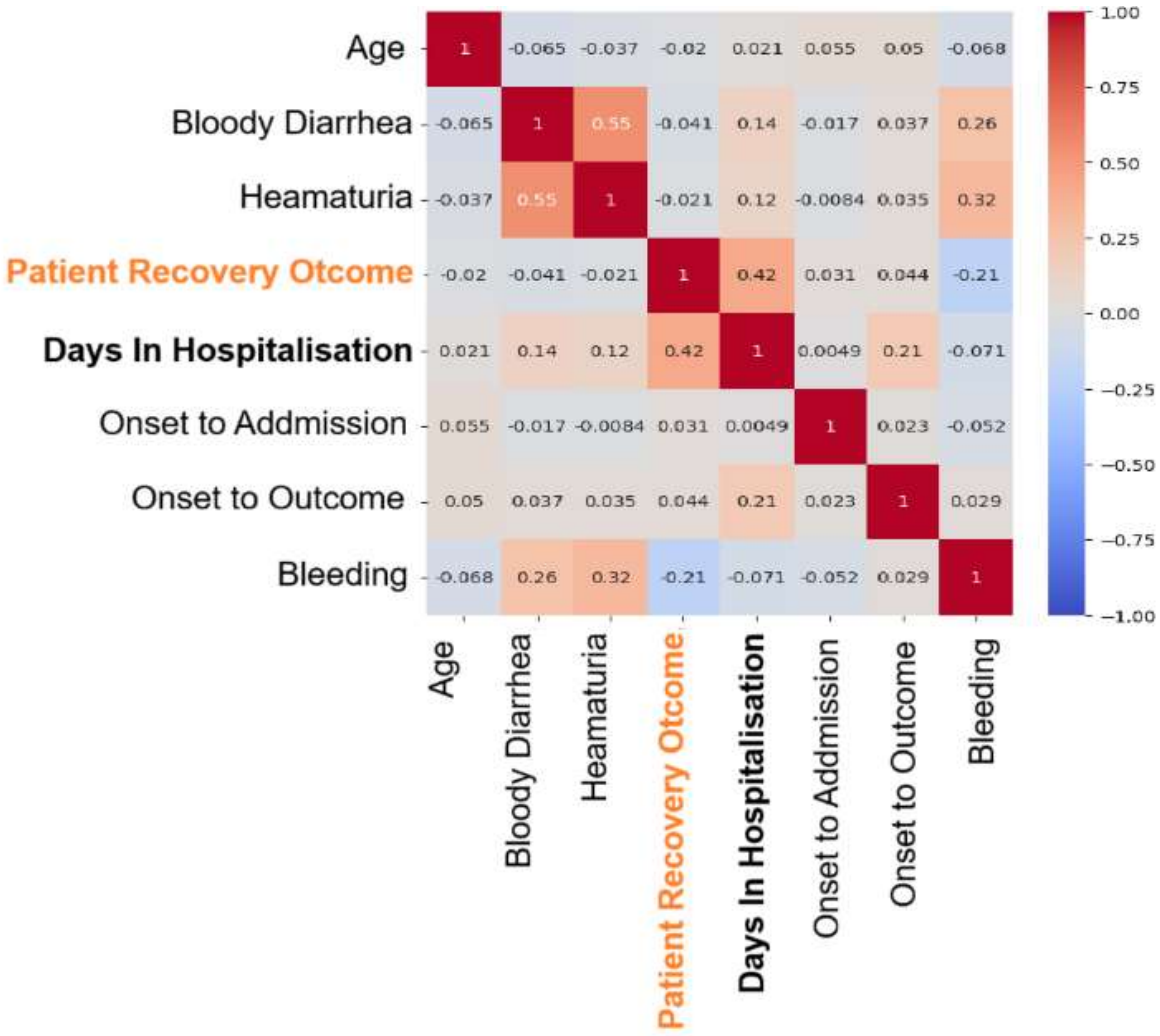

Figure 13 The correlation matrix of the features that are selected for this study.

The analysis of the feature importance from the Random Forest model, as shown in Figure 14, provides valuable information on the factors influencing outcomes in (CCHF) cases. Among the features, "Days in Hospitalization" holds the highest importance, highlighting its significant role in patient recovery chances. Both "Onset to Outcome" and "Age" are also highly influential, showing that the time between symptom onset and the final outcomes, as well as the patient's age, can have high impact on the patient outcome. Furthermore, "Onset to Admission" highlights the significance of early medical intervention as delays in seeking treatment can lead to poorer outcomes. Clinical symptoms like "Bleeding" moderately contribute to predictions, consistent with the severe manifestations of CCHF. Meanwhile, features such as "Bloody Diarrhea" and "Hematuria" have lower importance, suggesting that they are less decisive in differentiating outcomes. These findings are clinically relevant, guiding healthcare providers to prioritize key factors in diagnosis and treatment while informing public health interventions to encourage early detection and timely care. The analysis further underscores the role of specific variables that need more investigation (e.g., length of hospital stay and therapy effectiveness

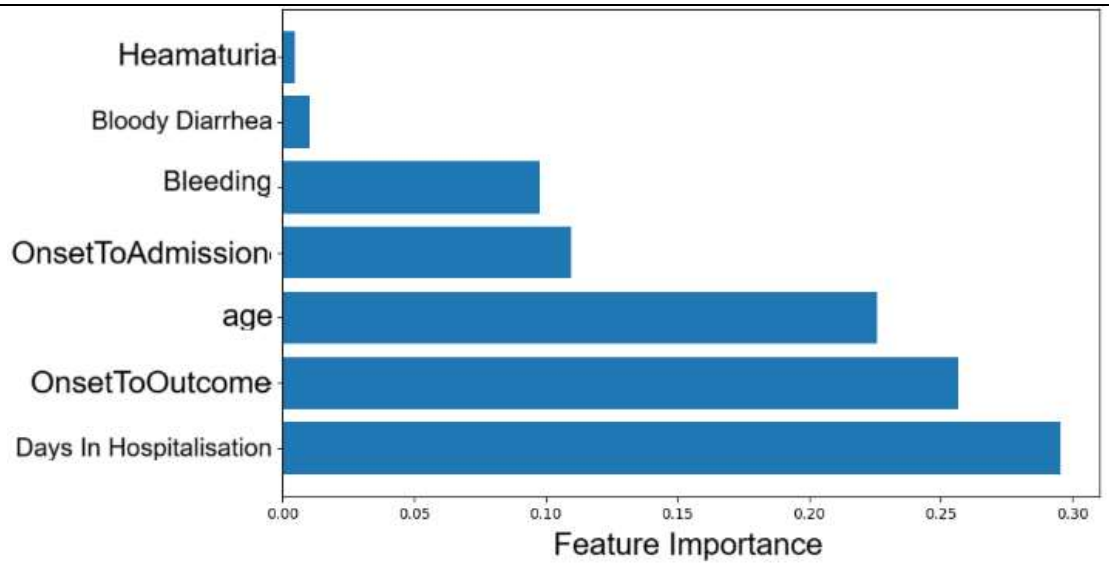

Figure 14: Optimized RF feature importance.

For further analysis of the impact of the risk factors, feature distribution across the risk level is shown in Figure 15. Three risk levels are considered. The red area indicates a high-risk level, the green area indicates a low-risk level, and the blue area indicates a middle-risk level. The range of values around the peak of each curve is the values that help to understand how the feature contributes to each risk level. The colors appear brown or dark brown when layering happens during intersection of the risk ranges.

- In Figure 15(a), it is obvious that younger adults (roughly 20-40 years old) are most likely to be classified as low risk (original colored blue but appears dark brown), while individuals in the middle age range (roughly 40-60 years old) show a higher prevalence of medium risk. Older adults (above 60 years old) show a slightly higher likelihood of being classified as high risk compared to younger individuals.
  However, the overall density of high risk (red area) remains the lowest. This aligns with many health conditions, where the risk might be lower in younger adulthood and is positively proportional to increasing age.
- Non-bleeding patients (Green area in Figure 15(b)) are predicted to be low risk, but bleeding appears to contribute to a non-linear relationship to risk.
- Figure 15(c) and Figure 15(e) show that patients without hematuria or diarrhea (value 0) are strongly predicted to be low-risk. However, having hematuria or diarrhea does not strongly indicate high-risk patients.
- Figures 15(d), (f), and (g) suggest that the most critical time period is around 5 days. Patients who exceed the first 5 days are likely to recover.

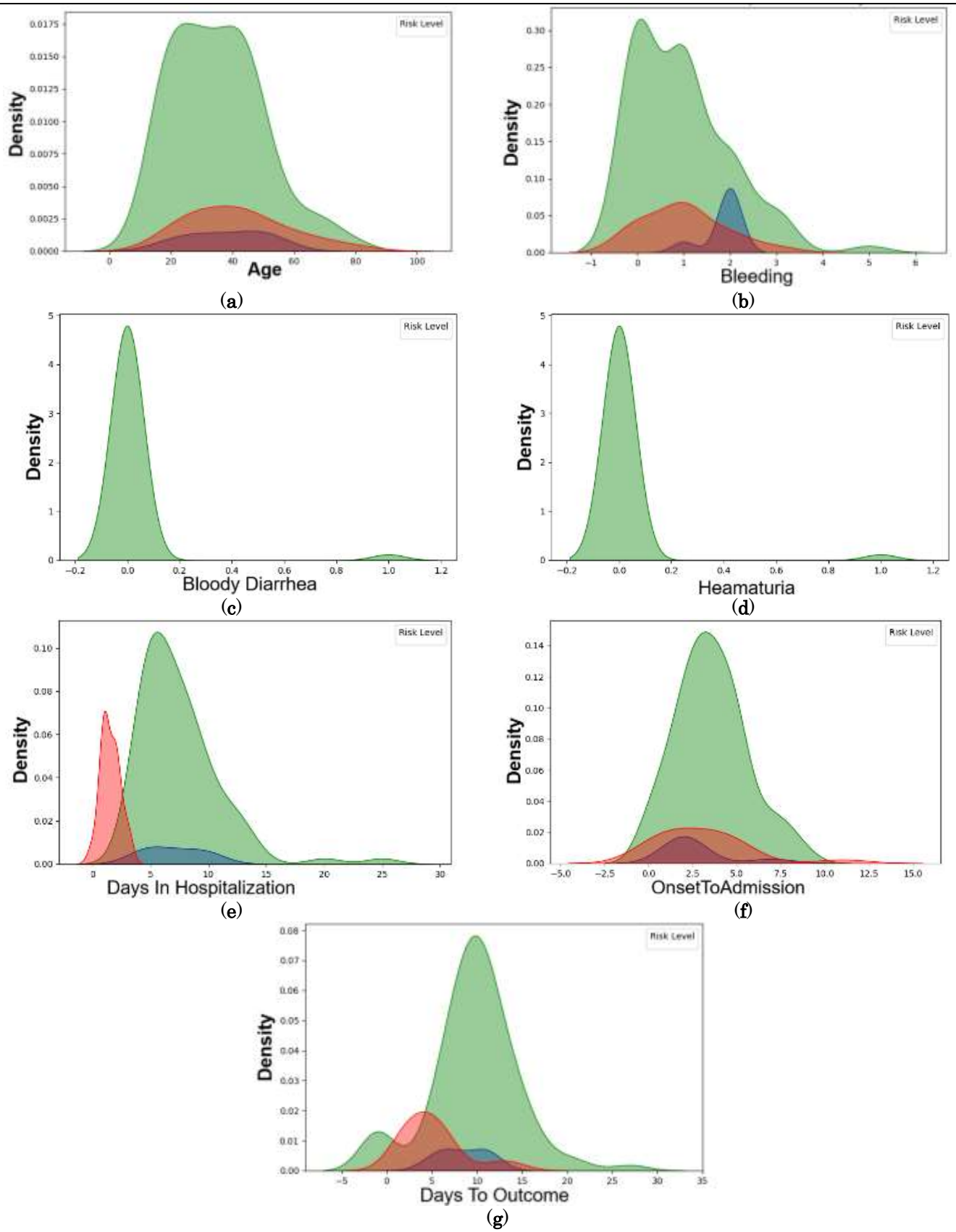

Figure 15: Features distribution by risk level. (a) Age, (b)Bleeding, (c) Hematuria, (d) Diarrhea, (e) Days in Hospital, (f) Days from Onset to Submission, (g) Days from Onset to Outcome.

5. Severity Prediction Application for the Health Care Provider

To assess the health specialists to manage and provide the best services for the patients, we built a web-based application that predicts the severity degree of the incoming patient. Figure 16 illustrates the web-based forms for entering the patient symptoms (in both English language web page (a, c) and Arabic language web page (b, d)). Then the degree of severity is predicted as either high-risk factors or not high-risk factors. For the application use, we built a new model that excluded the "Onset to Outcome" and "Days in Hospitalization" features, this is because it will not be provided when the patient first enters the hospital. Figure 17 illustrates the efficiency of the tested RF and DT models in terms of accuracy and Recall. RF1 and DT1 are the models trained without days in hospitalization, and without days to outcome features. The efficiency dropped after excluding these two features to reach 80% with a recall of 31% for the RF1 model, and accuracy of 74% and recall of 56% for the DT1 model.

Figure 16: The web application has 4 pages for CCHF prediction in English and Arabic.

To improve the efficiency, the SMOTEEN balancing method is applied that combines the oversampling method SMOTE (Synthetic Minority Over-sampling Technique) with the under-sampling method ENN (Edited Nearest Neighbors). RF2, and DF2 are the models trained on balanced dataset. The balancing resulted an improvement in the recall value to reach 50% for RF2, and 69% for DT2. On the other hand, the accuracy dropped even more to reach 60% for RF2 and 64% for DT2. The total improvement in recall is 19% which means reducing the false negatives. This is very important in healthcare providers' decisions. For further improvement, the bleeding features are disassembled to consider each bleeding type as a separate feature. The recall of RF3 reached 62% while the DT3 recall improved to 88%. The accuracy of RF3 reached 56% for RF3 and 60% for DT3. This suggests that having a balanced dataset with bleeding features combined as the sum of bleeding types, in addition to having each bleeding type also

contributing to the decision, produces the most accountable model to be applied. Thus DT3 model is considered for the web application.

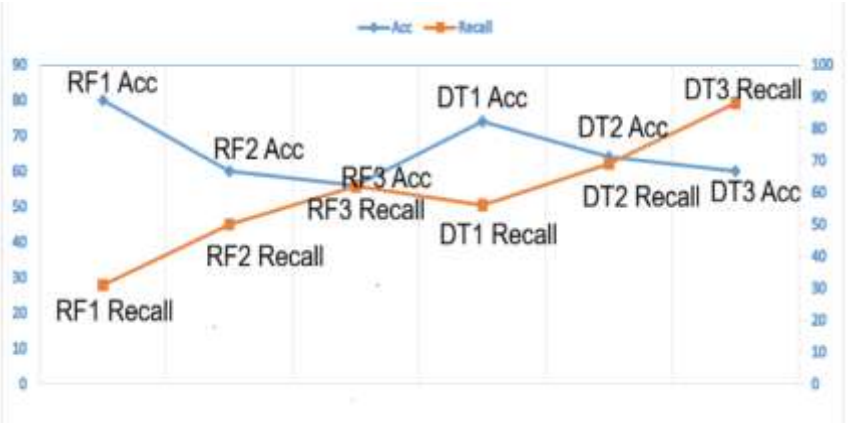

Figure 17: The Accuracy and Recall values of the six trained prediction models after removing the days in hospitalization and OnsetToOutcome features.

Figure 18 shows that the most important feature after the age factor is how quickly the patient enters the hospital to get the required health care. This result aligns with the results of the study to assess the risk factor in Section 4, which suggests the. importance of days of hospitalization as the most important factor.

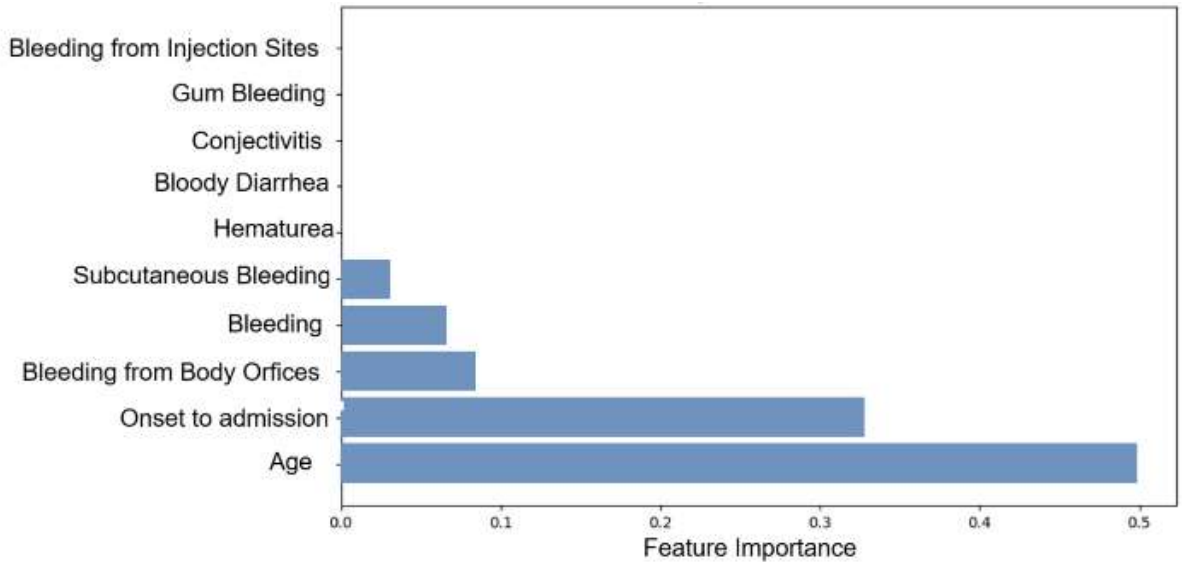

Figure 18: Feature importance excluding the days in hospitalization, and the onset to outcome information.

Although this research addresses an important literature gap, the study highlights that there are limitations, particularly in data availability and quality which could compromise the reliability of models. Generalization across other areas, as well as the complexity of the model, may impede its use in clinical practice. Moreover, the dynamic nature of CCHF epidemiology is probably not fully expressed, and by concentrating on mortality risk, other relevant health outcomes may be excluded. The lack of external validation also restricts the assessment of the model's performance in real practice. Additionally, ethical thoughts on possible biases in AI algorithms are also indispensable. Addressing these limitations is essential to enhance the model's

utility and effectiveness in public health applications.

6. Conclusions

This study highlights significantly the ability of Explainable AI and machine learning models to predict the threat of mortality for Crimean-Congo hemorrhagic fever (CCHF) in Iraq. By integrating multiple ML models.

Our studies achieved a predictive accuracy of 89%., given the limited size of the dataset which caused the classification accuracy not to exceed 90%. These findings show that the duration of a patient's clinical life is the most significant issue influencing patient recovery outcome, indicating that Iraq's healthcare system is well-equipped to manage CCHF cases effectively. In addition, the research introduced an Internet-based software for assessing real-time mortal threats that could assist medical decision-making and public fitness responses. Conclusively, this study shows that explainable machine learning models can help health officials identify CCHF patients at higher risk of death early, allowing faster treatment and better use of medical resources. However by revealing which factors mostly affects the patient outcomes. The models also supports informed public health decisions and can be adopted to improve outbreak response in other affected regions. Expanding this model to assess threat elements for various infectious diseases using similar methodologies may contribute to a broader AI-driven epidemiological framework, thus by applying AI for illness outbreak predictions, mortality checks, and public health tracking, this research can play an essential role in pandemic preparedness and crisis response techniques.

Conflicts of Interest: The authors declare no conflict of interest.

Funding: No funding has been received for this work.

Acknowledgments: The authors would like to thank Al-Nahrain University for the support in writing this manuscript, and the Ministry of Health, Baghdad, Iraq for their help in providing and authorizing the publishing of the dataset.

References

[1] Ergonul, O.; "Crimean–Congo hemorrhagic fever virus: new outbreaks, new discoveries". Curr. Opin. Virol., 2 (2): 215–220, 2012.

[2] Mertz, G. J.; "Zoonoses: Infectious Diseases Transmissible From Animals to Humans, Fourth Edition". Clin. Infect. Dis., 63 (1): 148, 2011.

[3] Ergönül, Ö.; "Crimean-Congo haemorrhagic fever". Lancet Infect. Dis., 6 (4): 203–214, 2006.

[4] Rehman, K.; Bettani, M. A. K.; Veletzky, L.; Afridi, S.; Ramharter, M.; "Outbreak of Crimean-Congo haemorrhagic fever with atypical clinical presentation in the Karak District of Khyber Pakhtunkhwa, Pakistan". Infect. Dis. Poverty, 7 (1): 59-64, 2018.

[5] Sah, R.; Mohanty, A.; Mehta, V.; Chakraborty, S.; Chakraborty, C.; Dhama, K.; "Crimean-Congo haemorrhagic fever (CCHF) outbreak in Iraq: Currently emerging situation and mitigation strategies – Correspondence". Int. J. Surg., 106 (1743-9191): 106916, 2022.

[6] Khwarahm, N. R.; "Predicting the Spatial Distribution of Hyalomma ssp., Vector Ticks of Crimean–Congo Haemorrhagic Fever in Iraq". Sustain., 15 (18): 13669, 2023.

[7] Alhilfi, R. A.; Khaleel, H. A.; Raheem, B. M.; Mahdi, S. G.; Tabche, C.; Rawaf, S.; "Large outbreak of Crimean-Congo haemorrhagic fever in Iraq, 2022". IJID Reg., 6: 76–79, 2023.

[8] Verdonk, C.; Verdonk, F.; Dreyfus, G.; "How machine learning could be used in clinical practice during an epidemic". Crit. Care, 24 (1): 265 2020.

[9] Sharma, M.; Goel, A. K.; Singhal, P.; "Explainable AI Driven Applications for Patient Care and Treatment". In Explainable AI: Foundations, Methodologies and Applications, Mehta, M., Palade, V. and Chatterjee, I., Eds.; Springer International Publishing: Cham, Switzerland, 135–156, 2023.

[10] Ak, Ç.; Ergönül, Ö.; Gönen, M.; "A prospective prediction tool for understanding Crimean–Congo haemorrhagic fever dynamics in Turkey". Clin. Microbiol. Infect., 26 (1): 123.e1–123.e7, 2020.

[11] Wang, Z.; Yang, C.; Li, B.; Wu, H.; Xu, Z.; Feng, Z.; "Comparison of simulation and predictive efficacy for hemorrhagic fever with renal syndrome incidence in mainland China based on five time series models". Front. Public Health, 12: 1365942, 2024.

[12] Zhang, T.; Rabhi, F.; Chen, X.; Paik, H. young; MacIntyre, C. R.; "A machine learning-based universal outbreak risk prediction tool". Comput. Biol. Med., 169: 107876, 2024.

[13] Colubri, A.; Hartley, M.-A.; Siakor, M.; Wolfman, V.; Felix, A.; Sesay, T.; Shaffer, J. G.; Garry, R. F.; Grant, D. S.; Levine, A. C.; Sabeti, P. C.; "Machine-learning Prognostic

Models from the 2014-16 Ebola Outbreak: Data-harmonization Challenges, Validation Strategies, and mHealth Applications". EClinicalMedicine, 11: 54–64, 2019.

[14] Forna, A.; Dorigatti, I.; Nouvellet, P.; Donnelly, C. A.; "Comparison of machine learning methods for estimating case fatality ratios: An Ebola outbreak simulation study". PLoS One, 16 (9): e0257005, 2021.

[15] Colubri, A.; Silver, T.; Fradet, T.; Retzepi, K.; Fry, B.; Sabeti, P.; "Transforming Clinical Data into Actionable Prognosis Models: Machine-Learning Framework and Field-Deployable App to Predict Outcome of Ebola Patients". PLoS Negl. Trop. Dis., 10 (3): e0004549, 2016.

[16] Jehad, R.; Yousif, S. A.; "Fake news classification using random forest and decision tree (j48)". Al-Nahrain J. Sci., 23 (4): 49–55, 2020.

[17] Schapire, R. E.; "The boosting approach to machine learning: An overview". In Nonlinear Estim. Classif. 1st ed.; Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B., Eds.: Springer, New York, NY, USA, 149–171, 2003.

[18] Chen, T.; Guestrin, C.; "XGBoost: A Scalable Tree Boosting System". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016; ACM: New York, NY, USA, 785–794, 2016.

[19] Podgorelec, V.; Kokol, P.; Stiglic, B.; Rozman, I.; "Decision trees: an overview and their use in medicine". J. Med. Syst., 26: 445–463, 2002.

[20] Blockeel, H.; Devos, L.; Frénay, B.; Nanfack, G.; Nijssen, S.; "Decision trees: from efficient prediction to responsible AI". Front. Artif. Intell., 6: 1124553, 2023.

[21] Elkahwagy, D. M. A. S.; Kiriacos, C. J.; Mansour, M.; "Logistic regression and other statistical tools in diagnostic biomarker studies". Clin. Transl. Oncol., 26 (9): 2172–2180, 2024.

[22] Hu, L.-Y.; Huang, M.-W.; Ke, S.-W.; Tsai, C.-F.; "The distance function effect on k-nearest neighbor classification for medical datasets". Springerplus, 5: 1304, 2016.