JCSDA, Vol. 1, No. 1, 14–38
DOI: https://doi.org/10.69660/jcsda.01012402
ISSN 2959-6912

Feature Selection Methods for ICU Mortality Prediction Model

Girma N. Alemneh
Department of Software Engineering,
HPC and Big Data Analytics Center of Excellence,
Addis Ababa Science and Technology University, Ethiopia
girma.neshir@aastu.edu.et

Hirut B. Ashagrie
Department of Internal Medicine
St. Paul's Hospital Mil. Med. College
Addis Ababa, Ethiopia

Lemlem K. Tegegne
Department of Software Engineering,
HPC and Big Data Analytics Center of Excellence,
Addis Ababa Science and Technology University, Ethiopia

The goal of this research is to offer insightful information that can improve Ethiopia's intensive care unit (ICU) services. There is an increased risk of patients' death in Intensive Care Units (ICUs). This is because of several variables, including preexisting medical issues, lack of resources, and delayed decisions. Healthcare professionals can better prioritize their patients in need of intensive care, distribute resources more efficiently, and enhance patient outcomes by using predictive models to estimate ICU mortality. ICU data is collected from five Ethiopian public hospitals to develop a machine learning method for predicting ICU mortality. The data includes demographic features, vital signs, lab results, and discharge status of 10,798 ICU dataset records. We employed a range of feature selection techniques, such as filters, wrappers, and embedding methods, to identify the most crucial features for mortality prediction. We also compared the performance of two machine learning algorithms, Random Forest and Decision Tree. These models are trained using ICU data with features encompassing age, length of stay, temperature, neutrophil, Diagnosis (DX) condition, PH, and Lymphocite. These features are selected using Recursive Feature Elimination (RFE). Using a number of different evaluation methods, including accuracy (99.7%), precision (99.4%), recall (98.8%), F1 score (99.1%), and area under the curve (AUC) (99.3%), Random Forest performed better than Decision Tree. Based on our findings, we made recommendations for healthcare practitioners and policy makers. We also suggest key future research directions for researchers in the area.

Keywords: feature selection, Intensive Care Units(ICU), patient mortality, healthcare data, machine learning, prediction model

1. Introduction

The effectiveness of a country's healthcare system has a significant impact on its development. Economic progress cannot be attained without having access to high-quality healthcare services. A significant investment is required for the infrastructure and trained labor of the health centers. The goal of this project is to provide valuable insights to enhance the services of Intensive Care Unit services (ICU).

ICU stands for intensive care unit for patients under severe emergency conditions requiring a significant amount of healthcare staff support and specialized medical equipment such as ventilators, monitoring devices, feeding tubes, drains, and catheters. On the basis of the report in [1], Ethiopia is one of the low- and middle-income countries (LMICs) where health system professionals frequently do less than half of advised evidence-based care measures. More than 8 million deaths a year in low-income nations like Ethiopia are a result of poor health systems. Six trillion dollars in economic welfare are lost as a result of these deaths. Moreover, research indicates that almost one-third of individuals encounter inappropriate treatment, brief appointments, inadequate communication, or lengthy in outpatient clinics. The capacity of health systems should provide care for complex and emergent illnesses is further hampered by weak referral systems and lack of platform integration. This causes a significant delay in retrieving patient health record data, which raises the death toll. The study in [1] also demonstrates the unequal access to high-quality healthcare across different nations, with disadvantaged and underprivileged populations receiving less effective treatment and a worse user experience. Due to specific care settings, medical issues, and demographic variables, people in MLICs are also more susceptible to receiving poor healthcare services. It is advised that the government enhance the number of clinics and health service providers with strong performance in order to achieve the SDG2 healthcare target, "Ensure healthy lives and promote well-being for everyone at all ages" [2]. The reasons for choosing this area are multi-folds: (1) to investigate the factors which contribute to the death of patients in ICU, (2) to support healthcare staff by providing useful insights about the fatality level of patients at ICU using data analytics, this will help staffs to decide reallocation of ICU who need most, (3) to assist health administrator to have useful insights from data like Mortality of patients to manage staffing and resources intelligently, (4) to provide insights of the dominant factors related to mortality for risk and disease control management, (5) to develop health care digital innovation that will help the policy makers to improve the health care facilities and service delivery up-to expectation of patients and minimize life loss. Numerous research papers [7, 8, 9, 10, 11] conducted in ICU are not adequate to predict the mortality of patients in ICU in the context of Ethiopian healthcare. The intensive care unit or emergency room in healthcare centers poses a number of issues linked to resource constraints, time constraints, and the need for important judgments. Moreover, the factors which contribute to the mortality of patients at ICU are not well investigated. Most ICU/Emergency department patients face unexpected bleeding, low blood pressure, irregular heartbeat, hypertension and breathing difficulties. In other words, it might be difficult to give accurate treatment to critically sick patients in low-resource settings: lack of equipment, inappropriate allocation, poorly trained staff, and inadequate infrastructure. A more accurate prediction of mortality of patients who are admitted to the ICU/emergency section could be pre-arranged often requires only a short time of surveillance before being discharged. Regardless of the existence of ICUs, many patients still pass away due to a lack of equipment or skilled personnel to give them the proper treatment. Critical care is needed for a growing number of ailments, including cardiac disease, diabetes mellitus, and renal failure. Besides, ICUs are better equipped in tertiary and specialty hospitals than in rural public hospitals [19]. The general objective of this study is to examine the feature selection methods to create a model to predict patient mortality at an intensive care unit.

In Ethiopian healthcare systems, the beneficiaries of the mortality prediction model encompasses patients, investors in the health sector, researchers in health, healthcare professionals, healthcare research centers/institutes/hospitals/clinics/labs/teaching referral hospitals, Ethiopian Public Health Institute (EPHI), and Ministry of Health.

2. Related Works

Feature extraction methods for ICU patient mortality prediction based on traditional machine learning have been explored in several studies. One study compared the performance of machine learning models with and without feature selection[39]. Another study leveraged classical machine learning techniques to identify the minimum feature set that best informs first-day mortality, with Elastic Net being the best performing method [39]. Additionally, another study used four popular supervised machine learning algorithms (Decision Tree, Random Forest, K-Nearest Neighbors, and Logistic Regression) to predict ICU patient mortality, Random Forest achieved a maximum accuracy of 0.87 [40]. These studies highlight the importance of feature selection and the use of various machine learning algorithms for accurate mortality prediction in the ICU. The papers in [7, 8, 9, 10, 11] used machine learning to predict the mortality of patients in ICU and few of these studies identified the factors determining the mortality of patients. Other studies in [3, 4, 5] used cohort study which helps to understand the mortality of patients in ICU in the past in the selected time frame and selected health centers.

In general, the factors causing mortality at ICU might vary across regions and countries and also varies over time, also vary based on the availability of quality of health care service, budget and other macroeconomic factors. In general, causes of death can be grouped into three categories: communicable (infectious), non-communicable (chronic and hypertension) and injuries [23]. For example, COVID-19 is an infectious disease that causes a great number deaths across the globe. However, the nature of symptoms and the rate of mortality vary across countries. The fatality rate of COVID-19 patients admitted to ICU is mainly determined by demography (i.e. age, gender, etc.) and comorbidities (cardiovascular diseases, cancers, diabetes mellitus, and chronic lung diseases) and the availability of health infrastructure [18].

3. Materials and Methods

The sequence of key tasks needed to complete the investigation is represented by the research flow. It offers the benefit of specifying the research-related procedures to be followed. Figure 1 shows how the research flow in this study is divided into different phases.

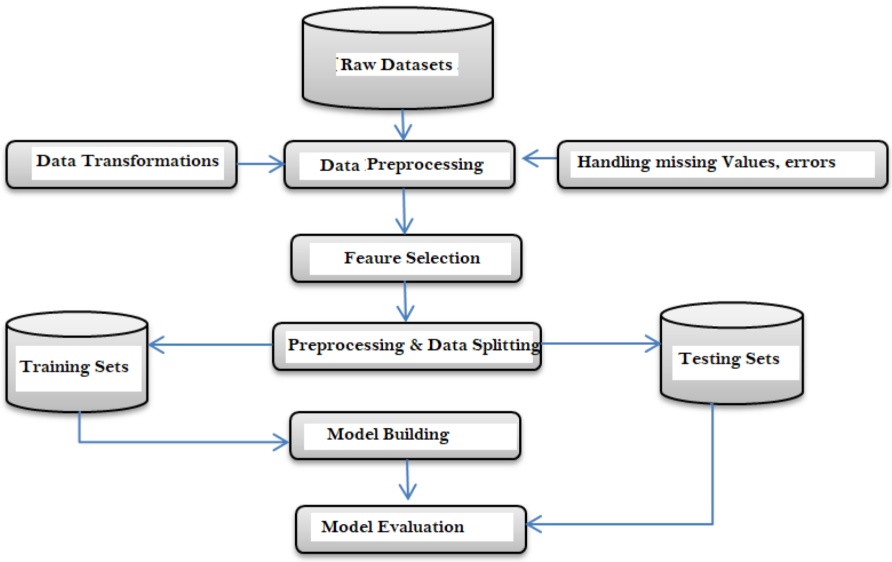

Fig. 1. Research Workflow.

3.1. *Data Sources, and Gathering Procedures*

Data Sources: This research utilized the population of ICU/emergency patients mainly focusing on non-communicable diseases. The data about ICU patients were collected from five hospitals, four of which are located within Addis Ababa, and one hospital located outside the city. The selection of hospitals was based on numerous criteria. Firstly, the four hospitals within Addis Ababa were chosen due to their diverse patient population and reputation for providing high-quality critical care services. These hospitals represent a range of healthcare settings, including public and teaching hospitals, capturing a broad spectrum of ICU patients in the city. In addition, a hospital from outside Addis Ababa is included to compare results and enhance the generalizability of the findings, allowing for a more comprehensive understanding of ICU care in different geographical locations. In addition, these hospitals have the capacity to provide comparable critical care services and willingness to participate in the study.

3.2. *Data Collection Procedures*

Data collection was conducted in accordance with established ethical guidelines and protocols. Besides the researchers' home institution, Institutional review board (IRB) approval was also obtained from each participating hospital to ensure patient privacy and data confidentiality. The data collection is carried out to capture a sufficient sample size and represent the typical ICU patient population. The data was primarily extracted from electronic health records (EHR) systems used by the participating hospitals. This method was chosen to ensure consistency and accuracy in data collection, minimize data entry errors, and facilitate the analysis of a large volume of patient information. The EHR systems contained comprehensive details on demographic features, vital signs, lab results, and discharge outcomes, allowing for a comprehensive analysis of ICU care. In addition, the collected data has been further validated and enhanced through iterative manual procedures to minimize data entry errors, missing values, and outliers. To ensure data accuracy and reliability, a thorough validation process was implemented. This involves cross-checking the extracted data against the original medical records to identify any discrepancies or missing information. Any inconsistencies were resolved through consultation with the healthcare providers and data entry personnel involved in the study. It is important to acknowledge the limitations of this methodology. Due to the retrospective nature of the study and reliance on EHR systems, there may be inherent limitations such as missing data, incomplete documentation, or variations in data quality across hospitals. Additionally, the inclusion of only five hospitals may not fully represent the diversity of ICU settings across the region or country. Despite these limitations, the chosen methodology allows for a comprehensive analysis of ICU data, providing valuable insights into the demographic features, vital signs, lab results, and discharge outcomes of ICU patients within and outside Addis Ababa. Providing a clear justification for the methodology used in data collection, will ensure the selected hospitals and data collection methods are appropriate for the research objectives.

3.3. *Data Preparation, Preprocessing and Transformation*

The majority of healthcare data is usually not kept in a digital format. To remedy such issues, the necessary data preparation processes must be undertaken. Data gathered from multiple hospitals and health institutions are encoded and combined to build a unified structure, and other activities were conducted at this stage. Data preprocessing is the activities carried out to improve data quality, which eliminates outliers, duplicates, irrelevant values, or handles missing values by adding appropriate defaults to missing values, encoding or removing irrelevant values and outliers from the data. Normalizing or standardizing the variables is also carried out in this stage. Generally, a variety of data preparation techniques have been utilized, relying on the purpose of the study and the type of data. Data transformation encompasses a variety of techniques depending on the nature of the data. Both categorical variables need to be transformed and encoded into numerical values and finally, all encoded and numerical values need to be normalized into standard scales.

3.4. *Feature Selection Methods*

In this stage, the most relevant features that are likely to impact ICU mortality are identified. In addition, domain expertise, literature review, or statistical techniques (e.g., correlation analysis, feature importance) are used to select the features. This step helps reduce noise, irrelevant features, reduce dimensionality in the data and in turn improves the accuracy of the mortality prediction model.

Feature selection methods aim to identify the most relevant subset of features from a larger set of input features. Filter methods are used to assess feature relevance based on intrinsic characteristics, while wrapper methods treat feature selection as a search problem by evaluating feature subsets using a specific learning algorithm. Embedded methods incorporate feature selection within the learning algorithm's training process. Each method offers unique benefits and considerations, such as dimensionality reduction, simplicity, incorporation of feature interactions, or algorithm-specific feature selection. The concepts of filter, wrapper, and embedded feature selection models are discussed below:

Filter Method: Without considering the performance of the learning algorithm into account, filter methods assess and rank features based on their inherent qualities. Statistical or metric-based measurements are used in filter approaches to evaluate feature importance without regard to particular learning algorithms. As each characteristic is evaluated separately, filter methods do not require an explicit search process. A subset of the highest-ranked features is chosen, and features are sorted according to their relevance scores as part of the filter feature selection process. The simplicity and computational efficiency of filter techniques are advantages. It is not dependent on the learning method either. Several filter techniques use statistical metrics to convey information about the significance of a feature. Conversely, filter approaches fail to take into account feature interactions and could miss significant feature combinations. Furthermore, it might not take into account the relationship with the target variable and is restricted to intrinsic qualities. Features that could be strongly linked or redundant could be chosen via the filter approach [37].

Algorithm 1: Filter Method: SelectKBest with Analysis of Variance(ANOVA)

Input:

- X: Feature matrix
- y: Target variable
- k: Desired number of features

Procedure: Input : X, y, k
Output: Selected features
Fit an ANOVA model on X and y to compute the F-value for each feature;
Rank the features based on their F-values in descending order;
Select the top-k features with the highest F-values;
Output: The set of selected features.

The pseudocode in algorithm 1 outlines the basic steps involved in the filter method using SelectKBest with ANOVA. It fits an ANOVA model on the feature matrix and target variable to compute the F-value for each feature. Then, it ranks the features based on their F-values and selects the top-k features with the highest F-values.

Wrapper Method: By analyzing the output of a specific learning algorithm, wrapper approaches evaluate feature subsets and approach feature selection as a search problem. By using a learning algorithm as a "black box," wrapper methods assess many feature subsets and choose the one that performs the best. By iteratively changing the feature subset in accordance with a search strategy and assessing the learning algorithm's effectiveness, wrapper techniques carry out an explicit search. The chosen feature subset and its matching performance on the selected learning algorithm are the outcomes of the wrapper feature selection process. The wrapper method assesses how feature interactions affect the effectiveness of the learning process. It offers adaptation to various problem kinds and flexibility to work with various learning algorithms. Wrapper approaches can identify the ideal feature subset to maximize the efficiency of the learning algorithm. However, because a learner is trained and evaluated iteratively, wrapper approaches are computationally costly. In addition, if the chosen feature subset doesn't adapt well to new data, there might be a risk of overfitting. Because the performance is so dependent on the particular learning method being utilized, it is not very generalizable [35].

Algorithm 2: Wrapper Method: Recursive Feature Elimination (RFE)

Input:
- X: Feature matrix
- y: Target variable
- *model*: Machine learning model to be evaluated
- k: Desired number of features

Procedure: **Input** **:** X, y, model, k
Output: Selected features
while
  the number of selected features is less than k
end
for each
  each feature f in remaining_features
end
Train the model using selected_features and f;
Evaluate the model on a validation set;
Select the feature with the highest performance metric as best_feature;
Add best_feature to selected_features;
Remove best_feature from remaining_features;
Output: The set of selected features.

The pseudocode in algorithm 2 outlines the basic steps involved in the RFE wrapper method for feature selection. It iteratively selects the best feature based on the performance of the model trained with the selected features. The process continues until the desired number of features (k) is reached.

Embedded Method: Feature selection is the most prominent phase in the training of the learning algorithm in embedded systems. By using regularization or built-in feature importance metrics during training, embedded approaches integrate feature selection within the learning process. Using properties unique to the learning algorithm, embedded techniques use feature selection as part of the training process, identifying and choosing features. The trained model and the features chosen based on their significance scores or coefficients are the outcomes of embedded feature selection [35, 37].

Journal of Computational Science & Data Analytics © AASTU Press

Algorithm 3: Embedded Method: L1 and L2 Regularization for DT/RF

Input:
- X: Feature matrix
- y: Target variable
- *model*: Decision Tree (DT) or Random Forest (RF) model
- k: Desired number of features

Procedure:
Input : X, y, model, k
Output: Selected features
Train the DT/RF model with L1 regularization on X and y;
Get L1 feature importance score from the DT/RF model based on the magnitude of the learned coefficients or feature importance measures (e.g., Gini importance for DT, feature importance for RF);
Sort features by L1 importance based on their L1 importance scores in descending order;
Select top-k features based on highest L1 importance scores;
Train the DT/RF model with L2 regularization on X and y;
Get L2 feature importance scores from the DT/RF model based on the magnitude of the learned coefficients or feature importance measures;
Sort features by L2 importance scores in descending order;
Select top-k features based on the highest L2 importance scores;
Output: The set of selected features.

In the pseudo-code in algorithm 3, the steps for training the machine learning model with L1 and L2 regularization remain the same. The difference lies in the method of obtaining feature importance scores. For Decision Trees (DT), you can use measures like Gini importance. For Random Forests (RF), you can use the feature importance scores provided by the RF model itself. The selected features are based on the importance scores obtained from the DT/RF model.

The embedded method produces more effective and efficient models by enabling immediate feature selection and learning. In order to find pertinent features, it can also take advantage of algorithm-specific properties. Embedded techniques, on the other hand, have limited transferability to various algorithms since they are closely linked with the selected learning algorithm. For complicated models, interpreting the feature relevance could be difficult. If the regularization is not correctly calibrated or the learning method is very sophisticated, there is a chance that it might overfit.

3.5. *Machine Learning Model Selection, Building and Evaluation*

The preprocessed data in earlier steps need to be divided into training and testing sets. The model is trained using the training set to discover different data attributes. Selecting the appropriate tools, techniques, and algorithm is one step in the model-building process. Learning from training data, generalizing existing knowledge, and utilizing untested data to develop predictions and achieve the goal were all essential steps in the model-building process.

3.6. *Evaluation of Models:*

Machine learning model evaluation is a crucial work that determines a model's efficiency and performance. The evaluation metrics encompassing accuracy, precision, recall, F-Score, Area Under the curve (AUC) and confusion matrix are commonly used to assess a proposed model's performance in testing machine learning prediction. These are described as follows:

Confusion Matrix: A confusion matrix is a table that summarizes the performance of a classification model by counting the number of true positive (TP), true negative(TN), false positive(FP), and false negative (FN) predictions. It provides a detailed view of the model's performance across different classes. The confusion matrix is typically represented as follows:

	Actual (+)	Actual (-)
Predicted (+)	TP	FP
Predicted (-)	FN	TN

The confusion matrix can be used to calculate various performance metrics such as accuracy, precision, recall, and F-score.

Accuracy(A): Accuracy measures the proportion of correctly classified instances out of the total number of instances in the dataset.

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

Precision (P): Precision is a metric that focuses on the accuracy of positive predictions.

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP + FP}} \tag{2}$$

Recall(R): Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive instances out of all actual positive instances in the dataset.

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP + FN}} \tag{3}$$

F-score: The F-score is a metric that combines precision and recall into a single value. It provides a balanced measure of the model's performance by considering both the ability to identify positive instances (recall) and the accuracy of positive predictions (precision).

$$\text{F-score} = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P + R}} \tag{4}$$

AUC (Area Under the Curve): AUC is a metric commonly used to evaluate the performance of binary classification models based on their receiver operating characteristic (ROC) curve. It measures the area under the curve, which represents the model's ability to distinguish between positive and negative instances. A higher AUC indicates better discrimination. A perfect classifier has an AUC of 1, while a random classifier has an AUC of 0.5.

3.7. *Software and Hardware Tools*

To create machine learning models, a variety of Python frameworks and tools including Scikit-Learn, Pandas, and Numpy are utilized. Python provides a huge library of solutions for a wide range of issues and uses. Furthermore, machine learning research requires hardware processing capabilities with rapid CPU/GPU speeds and big RAM capacities.

4. Experimentation, Results and Discussion

The study project starts with a succinct statistical analysis of the data and proceeds to a detailed description of the experimentation techniques and an assessment of the ones that have been performed. Lastly, conclusions on the model's functionality and feature selection methods are derived from the tests that were carried out.

Proposed Models: Random Forest (RF) and Decision Tree (DT) are two machine learning techniques that are used in the construction of the ICU patient mortality prediction model. After data preparation and relevant feature selection techniques are applied to the gathered ICU patient record data, the models are trained and assessed. By contrasting the model's performance using and not using feature selection techniques on the input dataset, the model is assessed. The most important procedures in dataset preparation involve dealing with missing values, categorical data oversight, and data cleansing. There are other vital phases involved as well. Following the preprocessing of the dataset, the preprocessed data is subjected to a variety of feature selection techniques in order to uncover significant numerical and/or categorical variables that enhance the ability to forecast the goal variable of mortality (of ICU patients). Following the application of feature selection approaches, the accuracy, precision, recall, f-measure, and confusion matrix were used to assess the suggested model.

Data Description: The data set for ICU patients entailing medical history are collected and organized in a manner that is appropriate for the construction of the ICU mortality predictive model. The dataset was collected from five public hospitals (Zewditu Referral Hospital, Ras Desta General Hospital, Yekatit 12 Hospital Medical College, Minilik II Referral Hospital and Debre Birhan Hakem Gizaw Hospital). The raw data on ICU patients was recorded in different hospitals during 2013-16. The collected ICU data has a total of 10,806 records. The researcher then went ahead and modified the dataset to meet the particular needs of the classification algorithms. After being collected from the five hospitals, the dataset was coded, arranged, and filtered from the patient information system into an Excel file. In order to better understand the data and plan future preparations, it is presently being explored. The data collected includes details about the medical histories of ICU patients, including diagnosis, vital signs, laboratory results, and demographics, in order to forecast the patient's prognosis. The collected raw data has incomplete information, missing values, outliers and inconsistencies. Some features encompassing patientId, patient name, phone number, MRN, user, nurse, physician, date of admission, discharge date etc. are removed. After preprocessing raw data, we obtained a cumulative count of 10798 ICU patient records, encompassing records with 23 distinct attributes, including one category feature (class), 16 numerical values, and 6 categorical values. As described in Table 1, the dataset comprises many variables, including Sex, Age, Address, Payment Type, Diagnosis Condition, Pulse Rate, Systolic Blood Pressure, Diastolic Blood Pressure, Body Temperature, Respiratory Rate, Oxygen Saturation, Pain Score, Random Blood Sugar, Oxygen Support, White Blood Cell Count, Neutrophil Count, Lymphocyte Count, Red Blood Cell Count, Hemoglobin, Platelet Count, pH Level, and Discharge Status.

Table 1. Description of features with its type (numerical or categorical)

Feature Name	**Description**	**Type**
Sex	Gender of the patient (Male/Female)	Categorical
Age	Age of the patient in years	Numerical
Address	Patient's residential address	Categorical
Payment Type	Method of payment for healthcare services	Categorical
Diagnosis(DX) Condition	Medical condition or diagnosis	Categorical
Pulse Rate(PR)	Heart rate in beats per minute	Num.(60-100)
Systolic Blood Pressure(SBP)	Highest pressure in the arteries during a heart beat	Num.(90-120)
Diastolic Blood Pressure(DBP)	Lowest pressure in the arteries between heart beats	Num.(60-80)
Body Temperature	Body temperature in degrees Celsius	Num.(36.5-37.5)
Respiratory Rate(RR)	Number of breaths per minute	Num.(12-20)
Oxygen Saturation(SpO2)	Percentage of oxygen saturation in the blood	Num.(95-100)
Pain Score (Pain)	Assessment of pain intensity (on a scale of 0-10)	Num.(0-10)
Random Blood Sugar(RBS)	Level of glucose in the blood	Num.(70-100)
Oxygen Support (Oxygen)	Type of oxygen support received	Categorical
White Blood Cell Count(WBC)	Number of white blood cells per microliter of blood	Num.(4,000-11,000)
Neutrophil Count	Number of neutrophils (a type of white blood cell)	Num.(1,500-8,000)
Lymphocyte Count	Number of lymphocytes (a type of white blood cell)	Num.(1,000-4,800)
Red Blood Cell Count (RBC)	Number of red blood cells in million per microliter of blood	Num.(4.5-5.5)
Hemoglobin (Hb)	Level of hemoglobin in the blood (12-16 g/dL for females, 13-17 g/dL for males)	Num.(12-17)
Platelet Count (Plt)	Number of platelets per microliter of blood	Num(150,000-450,000)
pH Level (pH)	Acidity or alkalinity level of a solution	Num.(pH 0-14)
Discharge Status (Discharge)	Status of patient's discharge from the hospital	Categorical
Length of Stay (LOS)	Duration of hospitalization in days	Numerical
Hospital	Name or code of the hospital	Categorical

After the data has been cleaned up and formatted properly for machine learning models such as RF and DT. The ICU dataset is scaled, encoded, and split into training and testing sets in order to build and assess the model Data Preprocessing: The dataset for the ICU mortality prediction algorithm used in this study was prepared using Python programming tools. The steps involved in data preprocessing encompass feature selection, standard scale transformation, encoding, and handling of missing values. To ensure processing for constructing ICU predictive models, every categorical data has been transformed into a corresponding numerical representation using the label encoder in the scikit-learn python library.

Dealing with missing values: Imputation techniques are used in this study to handle the missing values in the dataset. These techniques involve substituting computed or approximated values for the missing ones. Maintaining data integrity and ensuring the efficacy of machine learning models depend on the appropriate assignment of missing values. The SimpleImputer() function, a commonly used method for imputing missing values in both numerical and categorical variables, is utilized in this study from the Scikit-learn python library. The proposed methodology suggests imputing missing values in categorical features using the mode, which represents the most frequently occurring value. In the case of numerical variables, missing values are imputed with the mean value. By utilizing these imputation methods, this research aims to address missing data and enable reliable analysis and modeling within the machine learning framework.

Data transformation and encoding: Encoding and data transformation are essential steps in the machine learning preprocessing phase. In order to prepare the original input data for machine learning algorithm training, these techniques entail some manipulation. These procedures greatly enhance the efficiency and performance of the models by lowering noise, dealing with missing values, and formatting data so that algorithms can understand and analyze it. By following these basic steps, we can be sure that the data is properly prepared for the machine learning algorithm to extract useful insights and patterns.

Data imbalance handling: An imbalance in the target class's data set shows an obvious variation in the number of samples or occurrences across different categories. There are two commonly used techniques for addressing such imbalances: under sampling and oversampling. Reducing the number of instances in the majority class to match the number of occurrences in the minority class is termed as under sampling. On the other hand, over-sampling involves increasing the minority class's instance counts to equal that of the majority class. In this research, the target labeled dead is a majority class whereas improved is a minority class as depicted in Figure 2. It is illustrated that the dataset for the target category of discharge status is imbalanced. Imbalanced datasets can pose challenges in machine learning and statistical analysis tasks, as the model may be biased towards the majority class (i.e. Death) and may not perform well on the minority class (i.e. Improved). It is important to consider strategies such as resampling techniques, class weighting, or using evaluation metrics that account for imbalanced data when working with such datasets.

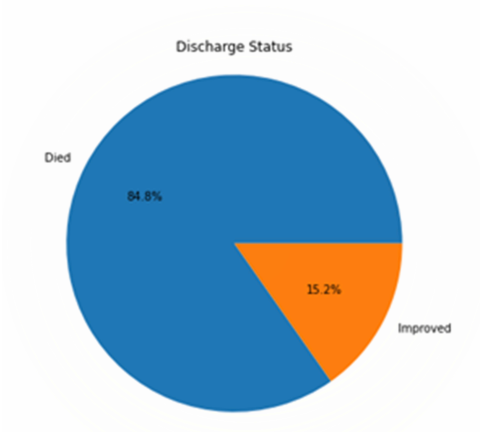

Fig. 2. Discharge Status.

In this study, we utilized the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a commonly employed approach in machine learning for oversampling. It generates synthetic samples from the existing dataset, thereby enhancing the classification capability, particularly for underrepresented classes. The SMOTE algorithm in [41] is applied to the datasets for handling imbalanced data. The original data set is split into two sets with an 80% training set and a 20% testing set. The training set before applying SMOTE has 8638 samples. This indicates that the original training dataset had 8,638 samples. The training set has been increased to 14644 samples by applying SMOTE. This indicates that SMOTE generated synthetic samples for the minority class to balance the dataset. In this case, the SMOTE algorithm has increased the number of samples to 14,644 in the training dataset. Testing set is used without applying SMOTE (i.e. 2160 samples). This suggests that the testing dataset consists of 2,160 samples and has not been modified by SMOTE. This means it seems that the SMOTE algorithm has been applied to address the class imbalance issue yielding an increased number of samples. In the training dataset, The testing dataset remains unchanged and is used for evaluation without applying any oversampling technique. The results obtained from SMOTE are used for the application of machine learning algorithms to predict the discharge status of ICU patients.

Feature Selection methods implementation: The implementation of feature selection encompasses various methods such as PCA, filter methods, wrapper methods, and embedded methods. In order to reduce the dimensionality of the feature sets, the filter selection method operates by evaluating the correlation between features and the target variable, thereby selecting relevant features. This approach incorporates the ANOVA test, implemented as f_classif from the Scikit-learn module. One of the wrapper feature selection methods is Recursive Feature Elimination

(RFE) which can be obtained from the Scikit-learn library. RFE selects features by recursively eliminating the least important ones based on model performance. In addition, embedded feature selection is implemented using L1(Lasso), L2(Ridge), and a built-in feature select technique with a random forest algorithm in the Scikit-learn library.

To put it in nutshell, though there are various feature selection methods, the choice of the above feature selection methods is stated as (1)SelectKBest with ANOVA is a widely used, computationally efficient, and interpretable method for ranking features based on their relationship with the target variable. (2)Recursive Feature Elimination (RFE) is a robust method that considers interdependencies among features and their relevance to the target variable. (3)Regularization such as L1 and L2 are an effective techniques for feature selection in high-dimensional datasets, mitigate overfitting, and improve model generalization

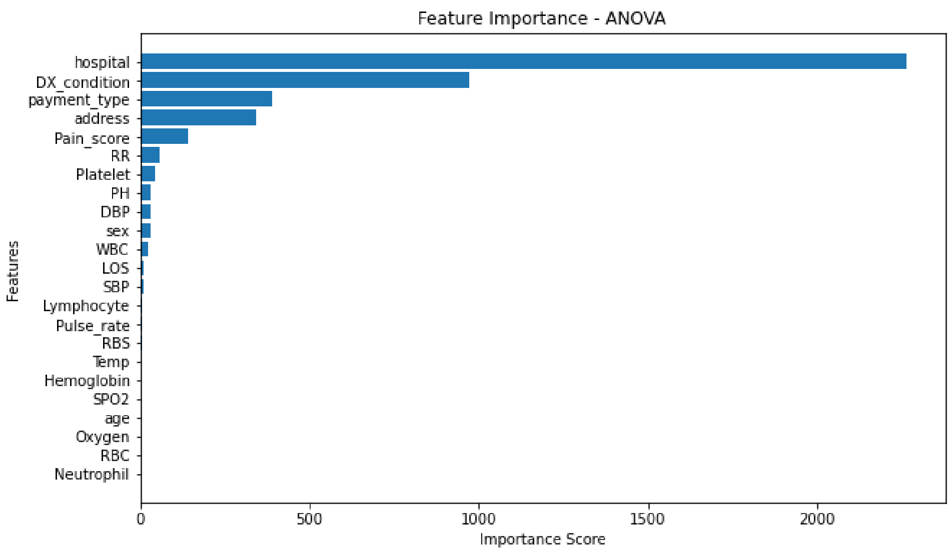

Fig. 3. Feasure Importance by Feature Filter methods (SelectKBest with ANOVA).

In Figure 3, feature importance scores are presented using two different methods: ANOVA.The analysis was conducted on 23 feature sets of ICU data. The results indicate that hospital, DX_condition, payment type, address, pain score, respiratory rate, and platelet are identified as the most important features by both feature selection methods. Conversely, the feature sets including neutrophil, RBC, oxygen, age, and SPO2 are determined to be the least important features.

In Figure 4, show feature importance by wrapper method (i.e. RFE feature selection). Age, length of stay, temperature, neutrophil, DX_condition, PH, Lymphocyte

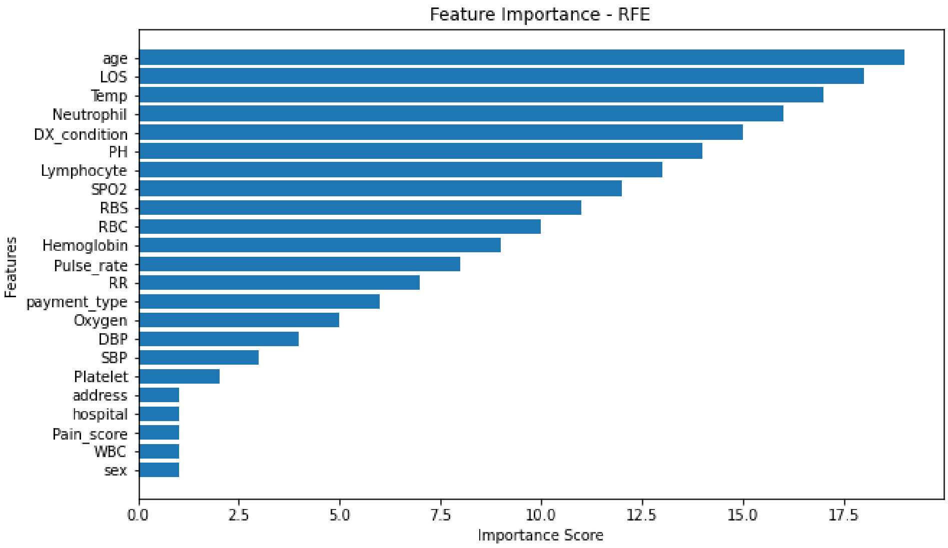

Fig. 4. Feature Importance by Feature Wrapper method using RFE.

are from top most important features using RFE. In Figure 7, feature importance is visualized using various methods of embedded feature selection. The subfigure 5 depicts the feature importance scores generated by L1 regularization. The subfigure 6 represents the feature importance obtained through L2 regularization techniques. These methods were utilized to assess the significance of features in the dataset. The results reveal that hospital, pain_score, DX_condition, pulse_rate, oxygen, temperature, respiratory rate, RBC, and age are identified as the most important features by both the RF and DT methods. On the other hand, pain score, SPB, Platelet, respiration rate, RBC, and RBS are determined to be the most important features based on both L1 and L2 regularization.

Implementation of Machine Learning Models: This research develops a predictive model using machine learning algorithms. The selection process prioritizes algorithms that demonstrate exceptional performance, as evaluated by their accuracy metrics.

After the data cleaning process is finished, the data set is divided into two subsets: a training set and a testing set. To enhance the training process, the fit() method is utilized to build the model.

The Decision Tree is a fundamental technique in machine learning that is commonly used for tasks involving classification and regression. This model is known for its simplicity and effectiveness as it utilizes a set of learned rules from the training data to guide decision-making. Decision trees are widely favored in various applications due to their easy comprehensibility, interpretability, and visualizability.

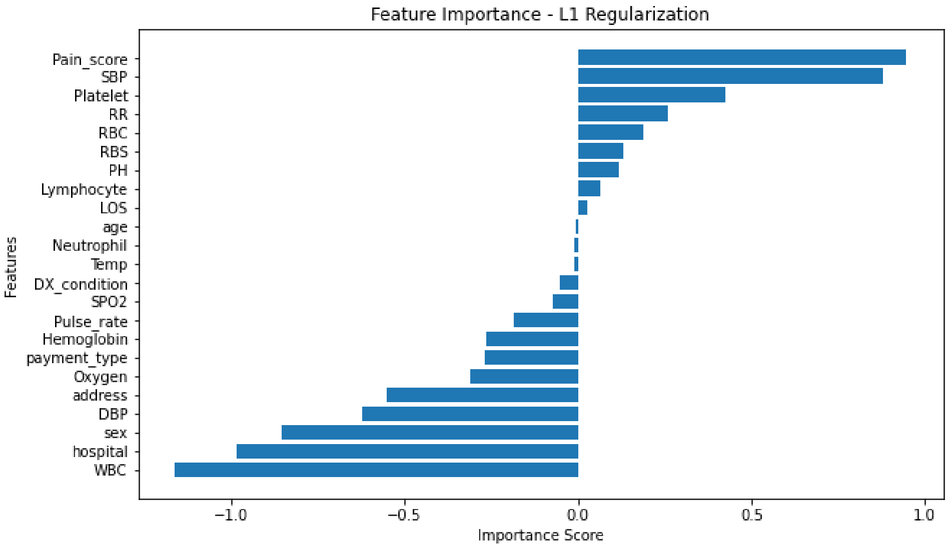

Fig. 5. Feature Importance by L1

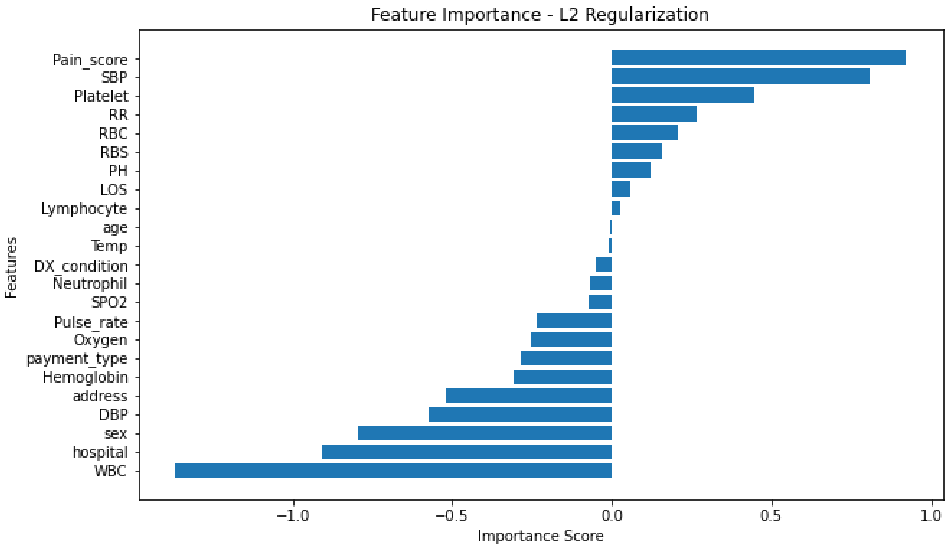

Fig. 6. Feature Importance by L2

Fig. 7. Feature Importance by Feature Embedded methods (L1 and L2).

In addressing the classification challenge at hand, the Decision Tree methodology is employed. This approach involves transforming the data set into a hierarchical structure that resembles a tree. This is achieved by organizing the feature values in a predetermined sequence. To implement this methodology, specific parameters and the DecisionTreeClassifier() library packages are utilized.

By utilizing Decision Trees, this research aims to tackle the classification task by creating a hierarchical structure that enables clear decision-making based on learned rules from the training data. The simplicity and interpretability of Decision Trees make them a valuable tool for understanding and analyzing complex datasets in a wide range of applications.

Random Forest (RF) is a popular ensemble learning technique in machine learning that is widely recognized for its effectiveness in both classification and regression tasks. This algorithm belongs to the decision tree family and is specifically designed to improve the predictive accuracy and robustness of individual decision trees. The implementation of Random Forest is carried out using the RandomForestClassifier() library packages, along with their respective parameters. By leveraging the power of multiple decision trees and their collective decision-making, Random Forest enhances the overall predictive performance by reducing overfitting and increasing generalization ability. Random Forest is highly regarded for its ability to handle complex datasets and produce reliable results. It combines the strengths of individual decision trees while mitigating their weaknesses, making it a valuable tool in various machine learning applications.

In summary, the justification for the choice of these predictive models is that (1) Random Forest is robust, scalable, handles high-dimensional datasets, and captures complex relationships between features and the target variable. On the other hand, the decision Tree is interpretable, captures non-linear relationships, and is widely used in healthcare research.

The preprocessed dataset has the total size of the dataset amounts to 10798 instances. Among the entire dataset, the number of patients falls into the following categories: Improved (1645), and Death (9153). The distribution of these classes in the dataset, consisting of two classes, is depicted in the graphical representation below.

Table 2. Models Performance

Model	Accuracy	Precision	Recall	F-Score	AUC	Confusion Matrix
RF	99.7	99.4	98.8	99.1	99.3	$\begin{bmatrix} 1829 & 2 \\ 4 & 325 \end{bmatrix}$
DT	99.5	98.5	98.5	98.5	99.1	$\begin{bmatrix} 1826 & 5 \\ 5 & 324 \end{bmatrix}$

Table 2 depicts a comparison of two different models, RF (Random Forest) and DT (Decision Tree), based on various evaluation metrics. The accuracy of the RF and DT models are 99.7 and 99.5, respectively, which means RF has a relatively better accuracy in predicting the outcome compared to DT. The RF and DT models have a precision of 99.4, and 98.5, respectively, which indicats that

RF has a high rate of correctly identifying positive cases. The RF and DT models have recall of 98.8 and 98.5 respectively, indicating that RF can identify a high proportion of positive cases compared to DT. The models such as RF and DT have F-score of 99.1 and 98.5, respectively indicating a good balance between precision and recall. The F-score indicated a similar balance. The RF model has an AUC of 99.3, suggesting a high level of discrimination between positive and negative cases. The DT model also has an AUC of 99.1, indicating comparable performance. The confusion matrix presents the number of true positives, true negatives, false positives, and false negatives showing breakdown of the model's predictions. The results indicated that the RF model has classified 1829 instances correctly as true negatives, 325 instances as true positives, 2 instances as false negatives, and 4 instances as false positives. Similarly, the DT model has categorized 1826 instances correctly as true negatives, 324 instances as true positives, 5 instances as false negatives, and 5 instances as false positives. In conclusion, both models look to perform well, with high accuracy, precision, recall, F-score, and AUC values. The RF model looks to have a slightly higher performance based on these metrics, but the difference is relatively small. It's important to take into account other factors such as computational complexity, interpretability, and specific requirements.

Outcome of Feature Selection: The dataset underwent Recursive Feature Elimination (RFE) for the purpose of feature selection. Since Recursive Feature Elimination (RFE) is a more robust method and considers interdependencies among features over the others. This procedure resulted from the identification of specific sets of features that can be utilized for training the Random Forest (RF) and Decision Tree (DT) models. These selected attributes are then utilized in a classification problem involving two distinct classes, namely "Improved" and "Death". The process of feature selection using the RFE method is used in the identification of the most relevant and discriminative features that contribute significantly to the classification task. By selecting the appropriate set of features, the models can accurately predict and effectively classify instances into their respective target classes.

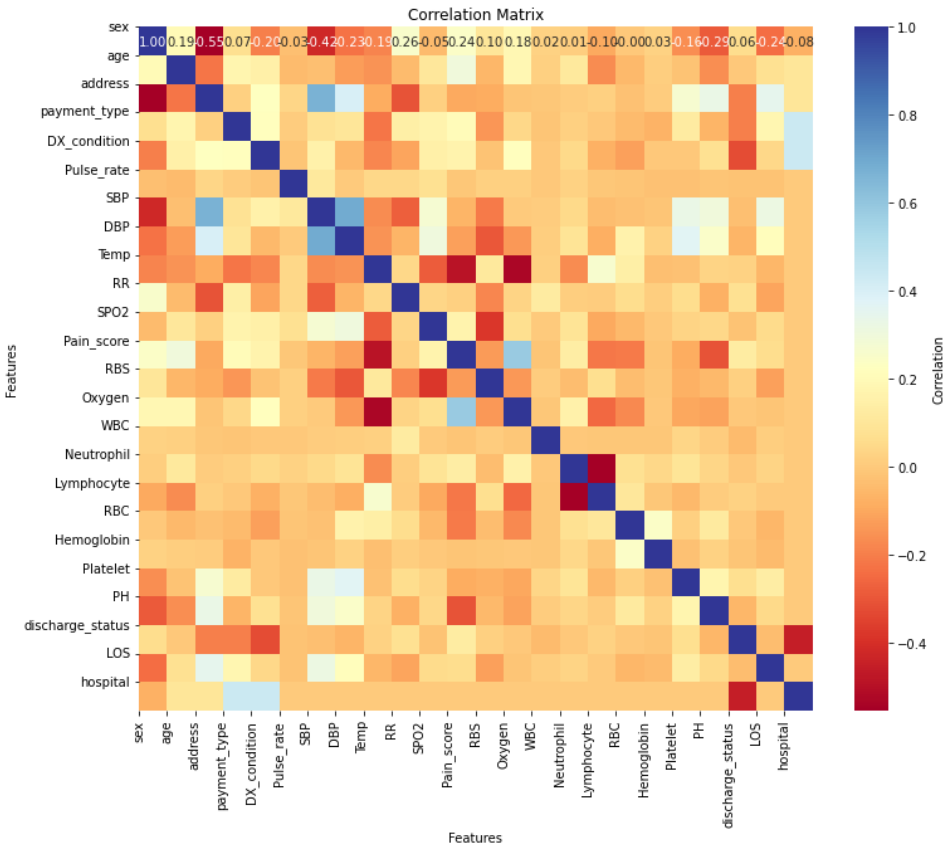

Fig. 8. Features heatmap based on Correlation method.

Figure 8 describes the dataset that includes twenty-four features of ICU patient records. It shows that there are several feature pairs that exhibit high correlation. Some of these pairs include SBP (Systolic Blood Pressure) and DBP (Diastolic Blood Pressure), DBP and address (which seems unusual), and oxygen levels and pain score. These feature combinations have a significant association, indicating that changes in one characteristic might be related to changes in the other. This information can be valuable for the classification task or analysis being performed on the ICU patient records. For instance, a high correlation between SBP and DBP indicates that changes in systolic blood pressure are likely to be accompanied by changes in diastolic blood pressure as well. This knowledge can aid in predicting or understanding the patient's condition. Similarly, a strong correlation between oxygen levels and pain score indicates that changes in one may be a sign of changes in the other. With the use of this data, the correlation between these variables can be investigated, and treatment plans or treatments can be developed in response to the findings. It's crucial to remember that a feature's association does not always imply a cause. Further analysis by domain experts is required to understand the underlying mechanisms and interpret the findings accurately.

5. Conclusions, Contributions and Recommendations

The study used data obtained from five Ethiopian public hospitals to develop a technique based on machine learning for predicting the mortality rate of ICU patients. The study uses several kinds of feature selection techniques, including filters, wrappers, and embedding algorithms, to determine which features are most important to the prediction objective. We used two machine learning algorithms' performances—Random Forest and Decision Tree—on the chosen attributes that were compared in the paper. The efficacy and dependability of both models for ICU patient mortality prediction have been demonstrated through the useful accuracy, precision, recall, F-score, and AUC values reported in the paper.

The research offers an in-depth look at the ICU data from Ethiopia, a low-income country with limited health resources and data availability, which could add to the body of knowledge on ICU patient mortality prediction. Through the use and evaluation of numerous feature selection techniques and algorithms for a difficult and demanding problem domain, the study also made a contribution to the field of machine learning. The research offered insightful information about the variables affecting the outcomes of ICU patients as well as the prospective applications of machine learning to assist in healthcare decision-making and resource allocation. The last but not least contribution is that creation of a dataset of 10,798 records for ICU patient mortality and length of stay prediction is a valuable resource for developing predictive models in critical care settings. Such models can have important implications for patient care, resource allocation, and healthcare planning.

Future research directions were suggested in the paper, which include expanding the areas of data collection to encompass more hospitals and regions, incorporating more variables and features pertaining to ICU patient care, investigating different machine learning models and techniques, and verifying the findings with additional data sources and the views of experts. The paper also recommended conducting further studies on the impact of machine learning on ICU patient mortality prediction on the quality of healthcare services, patient satisfaction, and health policy implications.

We recommend policy makers intervene in variables affecting the outcomes of ICU patients to consider in healthcare policy. Moreover, the applications of the mortality severity prediction model will assist significantly in healthcare decision-making and resource allocation and the dataset of 10,798 records for ICU patient mortality and length of stay prediction is a valuable resource for critical care settings. When more data is available, disease based model is more reliable and significant than the generic mortality model. In the end, we suggest explainable AI to be applied to provide further insights into the performance of the mortality severity model.

6. Ethical Considerations and Data Access Permissions

This research study is reviewed and ethically cleared by different bodies such as Addis Ababa Science and Technology University institutional ethical clearance board, Addis Ababa healh bureau, and participating hospitals are providing letters to allow access ICU data from well ensuring compliance with ethical standards. Access to the Intensive Care Unit (ICU) data was granted by the participating hospitals in this study. To protect patient privacy and confidentiality, all patient identifiers were anonymized prior to data analysis. The research team strictly adhered to data protection and privacy regulations throughout the study. The collected data was securely stored and accessible only to authorized personnel involved in the research project. Any personal information obtained during the study was handled with utmost confidentiality and used solely for research purposes.

7. Acknowledgment

The authors would like to express the acknowledgment of the funding support provided by Addis Ababa Science and Technology University. We are grateful for providing the financial support received from the Vice President's Office of Technology Transfer and Research, as well as the Research Directorate Office. We would also like to heartfully express our gratitude and appreciation to the HPC & Big Data Analytics Center of Excellence for their support and resources that made this research possible.

References

1. Kruk, M. E., Gage, A. D., Arsenault, C., Jordan, K., & et al. (2018). High-quality health systems in the Sustainable Development Goals era: time for a revolution. The Lancet Global Health Commission on High Quality Health Systems in the SDGs Era (HQSS), 6, e1196-e1252. https://doi.org/10.1016/S2214-109X(18)30386-3
2. WHO Regional Office for South East Asia. (2017). Monitoring the Health-Related Sustainable Development Goals (SDGs). SEARO, New Delhi, India.
3. Lalani, S., & et al. (2018). Intensive Care Outcomes and Mortality Prediction at a National Referral Hospital in Western Kenya. Annals of the American Thoracic Society, 15(11), 1336-1343. https://doi.org/10.1513/AnnalsATS.201801-051OC
4. Mekonnen Abate, S., & et al. (2021). Survival and predictors of mortality among patients admitted to the intensive care units in southern Ethiopia: A multi-center cohort study. Annals of Medicine and Surgery. https://doi.org/10.1016/j.amsu.2021.102318

5. Habtu, S. M. D. (2020). Cross Sectional Study on Mortality and Associated Factors in the Adult ICU of Myungsung Christian Medical Center, a Private Hospital in Addis Ababa, Ethiopia. Addis Ababa University. http://213.55.95.56/bitstream/handle/123456789/25273/Selam
6. Kifle, F., & et al. (2022). Intensive Care in Sub-Saharan Africa: A National Review of the Service Status in Ethiopia. Anesthesia & Analgesia. https://doi.org/10.1213/ANE.0000000000005799
7. Yun, K., Oh, J., Hong, T. H., & Kim, E. Y. (2021). Prediction of Mortality in Surgical Intensive Care Unit Patients Using Machine Learning Algorithms. Frontiers in Medicine, 8, 621861. https://doi.org/10.3389/fmed.2021.621861
8. Choi, M. H., & et al. (2022). Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. Scientific Reports, 12, 7180. https://doi.org/10.1038/s41598-022-11983-1
9. Choi, M. H., Kim, D., Choi, E. J., Jung, Y. J., Choi, Y. J., Cho, J. H., & Jeong, S. H. (2022). Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. Scientific Reports, 12(1), 7180. https://doi.org/10.1038/s41598-022-11226-4
10. Kong, G., Dong, X., Zhang, L., & Bai, X. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. BMC Medical Informatics and Decision Making, 20(1), 251. https://doi.org/10.1186/s12911-020-01271-2
11. Iwase, S., Nakada, T.-A., & Kawakami, E. (2022). Prediction algorithm for ICU mortality and length of stay using machine learning. Scientific Reports, 12(1), 12912. https://doi.org/10.1038/s41598-022-17091-5
12. Saadatmand, S., Ahmadi, M., Roozbeh, N., & Nasirian, H. (2022). Using machine learning in prediction of ICU admission, mortality, and length of stay in the early stage of admission of COVID-19 patients. Annals of Operations Research. https://doi.org/10.1007/s10479-022-04984-x
13. Subudhi, S., & Jena, S. S. (2021). Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. npj Digital Medicine, 4(1), 87. https://doi.org/10.1038/s41746-021-00456-x
14. Majhi, B. (2021). Mortality Prediction of ICU Patients Using Machine Learning Techniques. In Biomedical Data Mining for Information Retrieval: Methodologies, Techniques and Applications (pp. 1-20). Scrivener Publishing LLC.
15. Bitew, F. H., Tesfaye, D. J., Alemu, K., & others. (2020). Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey. Genus, 76, 37. https://doi.org/10.1186/s41118-020-00106-2
16. Ozyilmaz, A., Gokgoz, E., Kadioglu, H., & others. (2022). Socio-Economic, Demographic and Health Determinants of the COVID-19 Outbreak. Healthcare, 10, 748. https://doi.org/10.3390/healthcare10040748
17. Wotiye, A. B., & others. (2022). Factors Associated with ICU Mortality at

Hawassa University Comprehensive Specialized Hospital (HUCSH). Ethiopian Journal of Health Sciences, 32(3), 505. https://doi.org/10.4314/ejhs.v32i3

18. Sorci, G., & others. (2020). Explaining among-country variation in COVID-19 case fatality rate. Scientific Reports, 10, 18909. https://doi.org/10.1038/s41598-020-75848-2
19. Malelelo-Ndou, H., Ramathuba, D. U., & Netshisaulu, K. G. (2019). Challenges experienced by healthcare professionals working in resource-poor intensive care settings in the Limpopo province of South Africa. Curationis, 42(1), e1-e8. https://doi.org/10.4102/curationis.v42i1.1921
20. World Health Organization (WHO). (2020). The top 10 causes of death. Available online at: https://who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed: 9 December 2020.
21. World Health Organization (WHO). (2023). The top 10 causes of death. Available online at: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed: 13 January 2023.
22. Aryan, J., & Kumar, S. (2023). Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy. arXiv preprint arXiv:2305.05247. https://doi.org/10.48550/arXiv.2305.05247
23. Hernandez, M., Epelde, G., Beristain, A., Alberdi, A., Bamidis, P., & Konstantinidis, E. I. (2023). Effect of incorporating metadata to the generation of synthetic time series in a healthcare context. arXiv preprint arXiv:2305.05247. https://doi.org/10.1109/CBMS58004.2023.00341
24. Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review. (2023). Studies in Health Technology and Informatics. https://doi.org/10.3233/shti230063
25. Machine Learning for Synthetic Data Generation: A Review. (2023). https://doi.org/10.48550/arxiv.2302.04062
26. Gonzales, A. (2023). Synthetic data in health care: A narrative review. PLOS Digital Health. https://doi.org/10.1371/journal.pdig.0000082
27. Hanaya, R., Mohrova, Z., & Batech, M. (2023). A case for synthetic data in regulatory decision-making in Europe. Kliničeskaâ farmakologiâ i terapiâ. https://doi.org/10.1002/cpt.3001
28. Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review. (2023). https://doi.org/10.3233/SHTI230063
29. McDuff, D., Curran, T. R., & Kadambi, A. (2023). Synthetic Data in Healthcare. arXiv.org. https://doi.org/10.48550/arXiv.2304.03243
30. Xian, C. Z., de Souza, C. P. E., & Rodrigues, F. F. (2022). Health Outcome Predictive Modelling in Intensive Care Units. medRxiv. https://doi.org/10.1101/2022.12.15.22283527
31. Shah, D., Jariwala, D., Gupta, R., & Bharti, S. (2022). Machine Learning Based Hospital Mortality Prediction Using Synthetic Minority Oversampling Technique. https://doi.org/10.1109/iSSSC56467.2022.10051590

32. Ohnishi, Y., & Tenhunen, A. B. (2023). A New Risk Model Based on the Machine Learning Approach for Prediction of Mortality in the Respiratory Intensive Care Unit. Current Pharmaceutical Biotechnology. https://doi.org/10.2174/1389201024666230220103755
33. Fang, H. C. (2023). Machine-learning models for prediction of sepsis patients mortality. Medicina Intensiva (english Edition). https://doi.org/10.1016/j.medine.2022.06.024
34. Chi, T., & Wang, Y. (2023). Machine learning algorithm to predict the in-hospital mortality in critically ill patients with chronic kidney disease. Renal Failure. https://doi.org/10.1080/0886022x.2023.2212790
35. Chen, Z., Li, T., Guo, S., Zeng, D., & Wang, K. (2023). Machine Learning-based In-hospital Mortality Risk Prediction Tool for Intensive Care Unit Patients with Heart Failure. Frontiers in Cardiovascular Medicine. https://doi.org/10.3389/fcvm.2023.1119699
36. Gupta, S., Clarke, D. A. G., & Marín-García, D. (2023). Predicting Mortality Rate in ICU Using Machine Learning: A Study. Towards Excellence. https://doi.org/10.37867/te150112
37. Mohd Khalid, N. H., Ismail, A. R., Abdul Aziz, N., & Amir Hussin, A. A. (2023). Performance Comparison of Feature Selection Methods for Prediction in Medical Data. In Soft Computing in Data Science (pp. 1-12). Springer.
38. Chen, C.-W., Tsai, Y.-H., Chang, F.-R., & Lin, W.-C. (2020). Ensemble Feature Selection in Medical Datasets: Combining Filter, Wrapper, and Embedded Feature Selection Results. Expert Systems, 37, e12553. https://doi.org/10.1111/exsy.12553
39. Tasnim, N., & Al Mamun, S. (2023). Comparative Performance Analysis of Feature Selection for Mortality Prediction in ICU with Explainable Artificial Intelligence. https://doi.org/10.1109/ECCE57851.2023.10101553
40. Epifano, R. J., Ramachandran, R., Tripathi, A., & Rasool, G. (2023). A Comparison of Feature Selection Techniques for First-day Mortality Prediction in the ICU. https://doi.org/10.1109/ISCAS46773.2023.10182228
41. Chawla, N. V., & Bowyer, K. W. (2002). SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953