

Predictive Analytics of the Risk of In-Hospital Mortality of Heart Failure Patients in Intensive Care Unit

Umar Gambo Abubakar¹, Joshua Babatunde Agbogun², Temidayo Oluwatosin Omotehinwa³, Monday Jubrin Abdullahi⁴, Bethel Ugochukwu Nnamani⁵, Rosemary Ngozi Ariwa⁶, Daniel Gabriel Zhetu⁷ and Shedrach Yamlach Kamak⁸

^{1,2,5,6,7,8}Department of Mathematical Sciences, Taraba State University, P.M.B. 1167, Jalingo, Nigeria.

³Department of Mathematics and Computer Science, Federal University of Health Sciences, P.M.B. 145, Otukpo, Nigeria.

⁴Department of Computer Science, Faculty of Computing, Air Force Institute of Technology, P.M.B. 2104, Kaduna North, Kaduna, Nigeria.

maimuri2020@gmail.com¹, agbogunjoshua@tsuniversity.edu.ng², temidayo.omotehinwa@fuhso.edu.ng³, m.abdullahi@afit.edu.ng⁴, annamani.ugo@tsuniversity.edu.ng⁵, ariwarosemary@gmail.com⁶, dgzhetu@gmail.com⁷, kamak.ys@tsuniversity.edu.ng⁸

Abstract

Establishing and assessing appropriate predictive analytics model to forecast the likelihood of in-hospital death among heart failure patients admitted to ICU is the primary objective of this study. The research conducted a retrospective examination of health records obtained from MIMIC-III, a heterogeneous group of intensive care unit-admitted patients with heart failure. The research used machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boost (XGB), and Light Gradient Boost (LGB), to construct a prediction model using demographic data, vital signs, laboratory findings, and clinical history. To improve the accuracy and performance of the model, feature selection methods and cross-validation are employed. Predictive factors that contribute to the likelihood of in-hospital mortality among heart failure patients in a controlled environment of the ICU are anticipated to be identified in this research. Precision, Metrics such as Recall, F1-score, Area Under the Receiver Operating Characteristic Curve (AUROC), and K-fold Cross-validation were used to assess the performance of the model. This research aims to improve patient management outcomes, and makes valuable contribution to the expanding domain of predictive analytics in healthcare by enabling individualized and focused treatments, which eventually enhance the standard of care delivered to ICU patients suffering from heart failure. The potential of predictive analytics to improve clinical decision-making for patients with heart failure in intensive care units is highlighted in this work.

Keywords: Predictive Analytics, In-hospital Mortality, Heart Failure Patients, and the Intensive Care Unit.

1. Introduction

Heart Failure (HF) diseases is a clinical syndrome characterized by abnormalities in the structure and function of the heart (Bozkurt *et al.* 2021). Various history of machine learning in predictive analytics for healthcare reflects a continuous evolution driven by technological advancements, increased data availability, and a growing understanding of the complex relationships within healthcare datasets, as the technology continues to progress, the future is likely to bring even more sophisticated and personalized approaches to patient care (Mohd *et al.*, 2022). Predictive analytics is a data analysis technique that uses statistical algorithms and machine learning models to analyze historical data and make predictions about

future events or trends by identifying the patterns, trends, and insights that human analysts might skip (Smith, 2023). In the modern healthcare system, the ability to predict and positively manage adverse medical outcomes is of paramount importance. However, the current report on the outcome of the diagnosis in the healthcare system remains insufficiently addressed. Where the threat of heart failure remains the protracted medical syndrome, categorically defined as the incapability of the heart to pump blood effectively, resulting in critical health challenges. Conclusively, the study suggested that patients suffering from heart failure complications require intensive medical care attention, especially in the controlled environment of Intensive Care Units (ICUs) for serious clinical attention (Tetana, 2016).

Rabi *et al.*, (2024) developed a novel approach by incorporating Variational Auto-Encoder (VAE) approach into the Extreme Gradient Boosting Neural Network (XGBNet) techniques for predicting heart failure. They utilizes heart failure datasets from Kaggle, an open-source database, and achieved promising results with an accuracy of 92%, far better than previous prediction approaches, the findings indicates the effectiveness of integrating VAE and XGBoost in improving predictive performance. Predictive analytics plays a vital role in assessing the risk of in-hospital mortality for heart failure patients in the Intensive Care Unit (ICU) due to the increasing prevalence of heart failure menace (Jili *et al.* 2022). The study projected that by 2030, the prevalence of heart failure is expected to rise significantly, affecting millions of people worldwide. The observed in-hospital mortality rate for heart failure patients stands at approximately 9.97%, highlighting the urgent need for effective predictive models to enhance patient care and outcomes in ICUs (Jili *et al.* 2022). In-hospital mortality can be defined as the occurrence of patient death within the confines of a healthcare facility such as a hospital, clinic, or similar institution, while receiving medical treatment. The term in-hospital mortality was frequently used in medical research, epidemiology, and healthcare quality assessment to evaluate the efficiency of treatments, procedures, or medical interventions as well as understand the severity of the clinical illness in healthcare facility settings (Mohan, 2023). Heart Failure simply means that an individual heart can't pump the required oxygen-rich blood needed in the body. It does not mean the heart has stopped functioning; rather, blood is inadequately flowing in the circulation which possibly causes serious harm (Jacob *et al.*, 2017). Li *et al.* (2021), defines Heart Failure (HF) as the condition where the heart fails to perform adequately supply blood effectively due to chronic disease infections.

According to Raphael *et al.* (2022) the Intensive Care Unit (ICU) serves as a crucial life-saving entity, particularly for critically ill heart failure patients. They emphasize that heart failure patients present the highest-risk, most complex, and resource-intensive disease states. Similarly, Veli-Pekka *et al.* (2016) highlight that the clinical evaluation and management of cardiovascular conditions, including acute heart failure, in the ICU entail pharmacological and mechanical interventions. Almerund *et al.* (2007) suggested that the life-threatening nature of patients in Intensive Care Units (ICUs) requires skilled professional assistance as well as rigorous monitoring of the patient's acute signs using high-tech and sophisticated equipment. Ponikowski *et al.* (2016) their work earmarked that the larger proportion of heart failure patients may require sophisticated and high-tech machines should be mandated in the ICU for immediate intervention, and a functional Intensive Care Unit to be essentially equipped with professional and qualified medical workers including physicians and nurses in a doctor-to-patient ratio(1:1). Machine learning technology is significantly enhancing, transforming decision-making in various fields of human endeavors by improving computers through knowledge-based and utilizing online data and low-cost computations (Jordan & Mitchell, 2015). Big data and machine learning hold great promise for personalized medicine and effective public healthcare management, potentially improving population health and patient outcomes, reducing healthcare costs, and increasing the efficiency of healthcare delivery. Machine learning and predictive analytics have been used in healthcare for decades, in recent years, there has been a significant increase in the use of machine learning and predictive analytics in healthcare (Weintraub *et al.*, 2018; Akpan *et al.*, 2022; Omotehinwa *et al.*, 2023).

2. Literature Review

By measuring patient severity in the intensive care unit and developing a mortality-predicting model, Yang *et al.* (2021) sought to enhance patient outcomes. Their research work endeavored to develop an appropriate prediction model for intensive care unit (ICU) admissions by comparing six machine learning algorithms to the Acute Physiology and Chronic Health Evaluation (APACHE-II) score system-based prediction model. To prevent bias in the performance evaluation of the models, the research constructed the models using cohort data from the Medical Information Mart of Intensive Care-III (MIMIC-III) database. Performance estimation of the models was then accomplished by k-fold cross-validation. TreeSHAP was used to elucidate the interrelationships among variables in the XGB model to ascertain the result of the mortality prediction. Greater than APACHE-II (0.742) in accuracy, extreme gradient boosting achieved an accuracy of 0.858, surpassing the performance of the other models. In addition, with an AUROC of 0.79 and the following assessment metrics—accuracy, precision, recall, and F1—XGBM demonstrated superior calibration and discrimination compared to other algorithms (0.866, 0.853, 0.870 and 0.845).

In the research conducted by Guilan *et al.* (2020) models based on machine learning were created and verified for patients with sepsis who were at an elevated risk of dying in the hospital. Data obtained from the public database Medical Information Mart of Intensive Care-III (MIMIC-III) was used. Predictive models were constructed using four algorithms: Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), Gradient Boosting Machine (GBM), and conventional Logistic Regression (LR). A comparison was made between the models and the Simplify Acute Physiology Score of two (SAPS-II). In addition, assessment measures like AUROC, Bier scores, sensitivity, specificity, and calibration plot were implemented. The AUROC values for the LASSO, RF, GBM, and LR models are (0.829), 0.829, 0.845, and 0.833, respectively, whereas the SAPS-II value was (0.77). The research findings indicated that GBM exhibited superior prediction performance compared to the SAPS-II, surpassing all other models (LASSO, RF, and LR). Notably, machine learning-based models demonstrated even better performance. The research determined that the GBM model offered the most accurate and appropriate prediction of in-hospital mortality risk, as shown by an AUROC of (0.845).

The research conducted by Naghmeh *et al.* (2021) generalizes the cohort medical data of critically sick ICU patients. In order to forecast and investigate the critical determinants of mortality in the intensive care unit (ICU), the researchers conducted a comparative analysis of eight machine learning approaches: Decision Tree (DT), Neural Network (NN), K-nearest Network (KNN), Random Forest (RF), Linear Regression (LR), Adaboost, Bayesian Booting, Vote (DT+K-NN), and Vote (DT+K-NN+LR). Adaboost outperforms all other models chosen for high performance calibration (AUC = 0.966), and the model outcomes were successful in identifying and predicting the risk variables that influence mortality in the intensive care unit. The research included assessment criteria including AUC, recall, kappa, and F1, yielding the following results: 0.966%, 0.8788%, 0.8598%, and 0.8928%, respectively.

In their study, Li *et al.* (2021) conducted a retrospective cohort analysis compared machine-learning algorithms: Least Absolute Shrinkage and Selection Operator (LASSO), Extreme Gradient Boosting (XGB), and Logistic Regression with Get With The Guidelines-Heart Failure (GWTG-HF). These algorithms were employed to identify independent risk factors in both validation and derivation samples. The researchers utilized the MIMIC-III database, a publicly available medical cohort database, to compile the retrospective cohort data used in the model. In contrast to the traditional scoring model, Get With The Guidelines-Heart Failure (GWTG-HF), both machine-learning models exhibited superior calibration and discriminating performance. The Area Under the Curve (AUC) accuracy in both the derivation and validation groups favored the XGB model [AUC = 0.8515 (95% CI 0.7749 to 0.9115) and 0.8646 (95% CI 0.797 to 0.9201) for XGB, and 0.8194 (95% CI 0.6849 to 0.9030) and 0.8194 (0.9201 to 0.9205) for LASSO]. XGB demonstrated superior discrimination and calibration, establishing it as a reliable tool for predicting in-hospital mortality.

In a comparable investigation, Cheng *et al.* (2023) built a machine-learning forecasting model by leveraging two publicly available databases and the Extreme Gradient Boost algorithm. The model utilized the MIMIC-IV dataset for training purposes (training set) and the eICU datasets for external validation (test set). Its objective was to ascertain the likelihood of in-hospital mortality for patients admitted to the intensive care unit with heart failure. In contrast to logistics regression and the established Get with the Guideline Heart Failure model (GWG-HF model), the XGB model demonstrated superior performance in terms of discrimination, calibration, and clinical net benefit. This was evidenced by the XGB model's Bier score of 0.10% and an area under the receiving operating curve of 0.771 (with a 95% confidence interval of 0.757-0.784). With the use of this approach, all causes of in-hospital mortality among heart failure patients admitted to the intensive care unit may be precisely evaluated and categorized.

A comprehensive predictive risk model for heart failure mortality was developed by Fen *et al.* (2018) in comparison to two ensemble survival functions, namely the improved Random Survival Forest (iRF) and the conventional Random Survival Forest (RSF). The study utilized the MIMIC-II dataset of heart failure patients. The research used the splits rule of the long-rank test, which maximizes the disparity between the survival daughter nodes on each survival tree in the traditional and modified random survival forest-based trees, respectively. The results were determined using out-of-bag C-statistics, which achieved an accuracy score of 0.804 and 0.821, respectively. As a result, the iRSF exhibited a higher level of efficacy in predicting the clinical mortality of patients with heart failure compared to the survival function.

The research conducted by Hsich *et al.* (2011) effectively identified the critical risk factors for survival in patients with systolic heart failure through the application of Random Survival Forest (RSF) and traditional Cox proportional hazard regression to 2231 adult patients with systolic heart failure undergoing various cardiopulmonary stress tests. The prediction outcomes indicate that the RSF and Cox regression models provide comparable results (out-of-bag C-index of 0.705 for vs 0.698 for Cox proportional hazard models). It was determined that the RSF model and the conventional Cox proportion provide medical practitioners with a more intuitive method for detecting significant risk factors associated with all-cause mortality.

Tian *et al.* (2023) examined 67 features of Heart Failure with mildly reduced Ejection Fraction (HFmrEF) patients cohort dataset extracted from Heart Failure Care Unit, Fuwai Hospital, Chinese Academy of Medical Science and Peking Union Medical College (CAMS & PUMC) was conducted using two feature selection techniques: feature correlation and clinical significance; and the CoxBoost algorithm with 10-fold cross validation applied to four machine learning algorithms: eXGB, RF, SVM, and logistic regression. The eXtreme Gradient Boost model (with an area under the receiver-operating characteristic curve of 0.916) demonstrated superior performance compared to the other three algorithms, with an accuracy of 85.5 percent and a precision of 70.35. (95 percent CI, 0.887-0.945). 0.12 was the Brier score.

Mathew *et al.* (2022) conducted research that examined the integration of Social Determinants of Health (SDoH) by developing a predictive model using data sourced from the Get With The Guidelines-Heart Failure (GWTG-HF) medical record. Random Forest and other machine-learning approaches were used to construct the Race-Specific and Race-Agnostic models, respectively. As a result of the models' enhanced discrimination and appropriate calibration (C-statistics 0.81 for black patients and 0.80 for white patients).

Adult patients admitted to the general wards of five hospitals within the preceding twelve months were included in the retrospective cohort study and Deep Learning-based Early Warning score (DEWS) on cardiac arrest patients conducted by Lee *et al.* (2021). The objective of this research was to assess the efficacy of the prediction alarming and timeless method in identifying patients at risk of in-hospital cardiac arrest (IHCA) mortality in comparison to the Modified Early Warning Score (MEWS). In terms of the internal and external AUROC scores, DEWS performed better than MEWS (0.860 and 0.754 against 0.905 and 0.785, respectively). The results of the research indicate that the DEWS outperformed the MEWS in

all three dimensions when it came to predicting, alerting, and enduring performance, as well as being a viable and efficient tool for rapid response systems and screening in-hospital cardiac arrest mortality.

Khera *et al.* (2021) conducted an assessment of machine learning models designed to forecast mortality subsequent to acute myocardial infarction. The data used in this investigation came from the American College of Cardiology Chest Pain-MI Registry. A neural network, a meta-classifier model, and extreme gradient descent boosting (XGB) were designed as machine learning models. On the basis of C-statistics, XGB and meta-classifier models exhibited more discriminatory power than logistic regression (0.90 for XGB).

Li *et al.* (2022) conducted a comparative analysis of four machine learning algorithms—XGB, RF, LR, and SVM develop an interpretable model to forecast the mortality risk for intensive care unit (ICU) patients with heart failure. The study utilized retrospective cohort data from the Collaborative Research Database eICU-RCD and employed the Shapley Addictive Explanation (SHAP) method to elucidate the extreme gradient boosting process. In contrast to alternative models, XGB demonstrated superior predictive performance, as evidenced by its AUC of 0.824 (95 percent CI 0.7766-0.8708). Consequently, it was deemed appropriate for implementation as an interpretable predictive model that would assist healthcare professionals in accurately forecasting the mortality risk associated with heart failure in the intensive care unit.

More so, researchers Dineo *et al.* (2022) analyzed 30 studies that had produced mortality prediction models for heart failure patients. They combed through datasets from several academic databases as well as the Google Scholar domain. Only 30 out of 260 publications were deemed suitable for the review in the research. Patients' ages ranged from 71 to 716,790, with a median of 72.1 years, among the study samples. Additionally, the AUC accuracy of the created models varies from 0.49 to 0.92 according to the papers that were examined. Hospitalization data for patients also have an accuracy range of 0.47 to 0.84 according to the AUC. In addition, the nine machine learning algorithms that were examined in the publications are as follows: logistic regression, random forest, decision trees, gradient boosting, neural networks, support vector machine, naïve Bayes, k-nearest neighbor, and Bayesian network. In addition, the study found that when it came to building and evaluating predictive models, around 63% of the publications used Logistic Regression (LR) approaches, 53% used Random Forest (RF) techniques, and 37% used Decision Tree (DT) strategies. Data from all across the globe, including North America, Europe, Asia, South America, Australia, and New Zealand, was utilized by all of the produced models. The examined studies did not include any patient data from African nations or the Middle East, according to this analysis. Additionally, the evaluated papers demonstrated that AUC is the most often used measure for evaluating model performance. Out of 30 publications, Nakajima's research had the highest AUC coded at 0.92, which was Random Forest.

3. Methodology

The MIMIC-III dataset obtained from [kaggle.com](https://www.kaggle.com) comprises 1177 entries and 51 columns, featuring a diverse range of demographic, clinical, and laboratory parameters related to heart failure patients in Intensive Care Units (ICUs). Each row corresponds to an individual patient, identified by a unique ID, and is categorized into different groups. Key features include patient outcomes (binary, float), age (int), gender (int), BMI (float), and various medical conditions such as hypertension, atrial fibrillation, coronary heart disease, diabetes, deficiency anemias, depression, hyperlipemia, renal failure, and Chronic Obstructive Pulmonary Disease (COPD) – all of which are represented as integer values. Additionally, the dataset encompasses vital signs such as heart rate, blood pressure, respiratory rate, temperature, and oxygen saturation, expressed as float values. Laboratory parameters include hematocrit, RBC indices, leukocyte count, platelet count, coagulation factors, cardiac biomarkers, electrolytes, and blood gas values, all

represented as float64. The dataset exhibits variations in missing values across columns, requiring careful consideration during analysis. The provided information, along with the inclusion of both numerical and categorical variables, forms a rich foundation for exploring and predicting in-hospital mortality among heart failure patients in ICUs.

3.1 Data Preprocessing

Data preprocessing is a crucial step in preparing raw datasets for analysis and model training. In the provided code, the process begins by loading the dataset into a Pandas DataFrame. The next step involves calculating the percentage of missing values for each column, and below is the graphical visualization of each column with percentage of missing values:

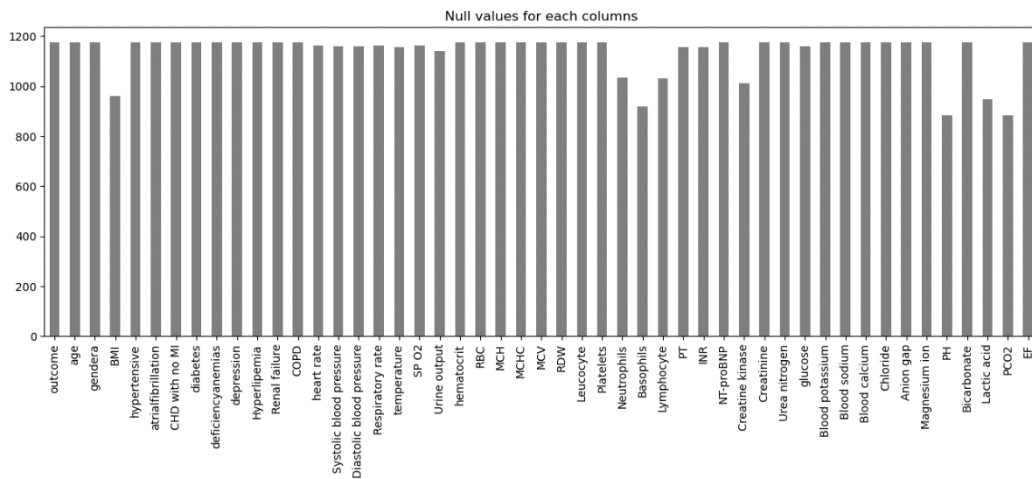


Figure 1: Percentage of missing value plot for each column

The dataset added to the Data Frame is preprocessed to handle missing values using SimpleImputer. This function used to data imputation in a DataFrame named df. It fills missing values in various columns with either the mean or mode of the respective column. Subsequently, all data types within 'data_cleaned' are converted to float64 to ensure uniformity. The selected features for modeling are then defined, focusing on relevant clinical and demographic parameters while excluding unnecessary columns like 'group,' 'ID,' and 'outcome.' The final dataset ('df') is a refined subset, containing only the specified columns. These preprocessing steps collectively contribute to a cleaner and more standardized dataset suitable for subsequent model training and analysis.

3.2 Exploratory Data Analysis

According to Brian, (1998), Exploratory Data Analysis (EDA) is a methodological approach used to gain a preliminary understanding of a dataset. It involves examining and visualizing the distributions and relationships between variables within the dataset. By doing so, EDA uncover patterns, trends, and anomalies that may exist in the data. He further described EDA not only helps in understanding the structure of the data but also assists in generating hypotheses and guiding the selection of features for further analysis. In essence, it serves as a crucial initial step in the data analysis process, providing insights that can inform subsequent analytical approaches and decision-making processes. This study employs truncated Singular Value Decomposition (SVD), t-distribution Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Protection (UMAP) to reduce the dimensionality of the data. Scatter plots are utilized to represent the data in a two-dimensional space, which facilitates the observation of patterns and

clusters and the comprehension of complex data structures. To facilitate precise exploratory data analysis and get insight into the fundamental structure of the data, the following feature categories are employed:

- Demographic information encompasses all non-clinical data pertaining to a patient, including age, gender, and Body Mass Index (BMI).

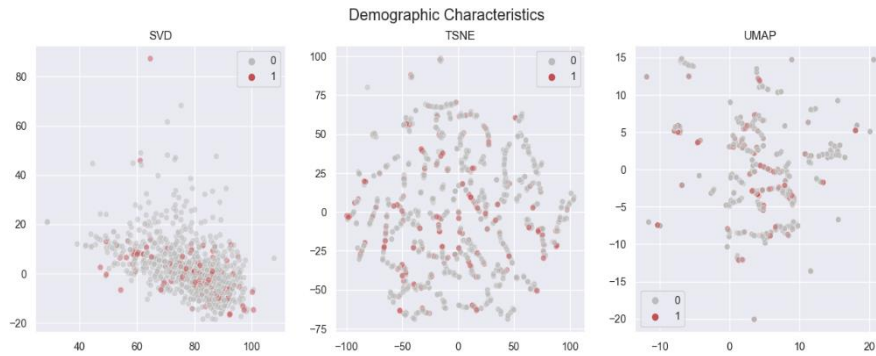


Figure 2: Demographic Characteristics

- Comorbidities refer to the presence of additional diseases in patients with heart failure, although not necessarily associated with heart complications. These conditions may include chronic kidney disease, chronic obstructive pulmonary disease, hypertension, atrial fibrillation, ischemic heart disease, diabetes mellitus, depression, hypoferric anemia, and hyperlipidemia (COPD).



Figure 3: Comorbidities Characteristics

- The following vital signs—Heart Rate (HR), systolic blood pressure (SBP), Diastolic Blood Pressure (DBP), mean blood pressure, respiratory rate, body temperature, Saturation Pulse Oxygen (SPO2), and urine output—are intrinsically linked to the health of the heartbeat: (first 24 h).

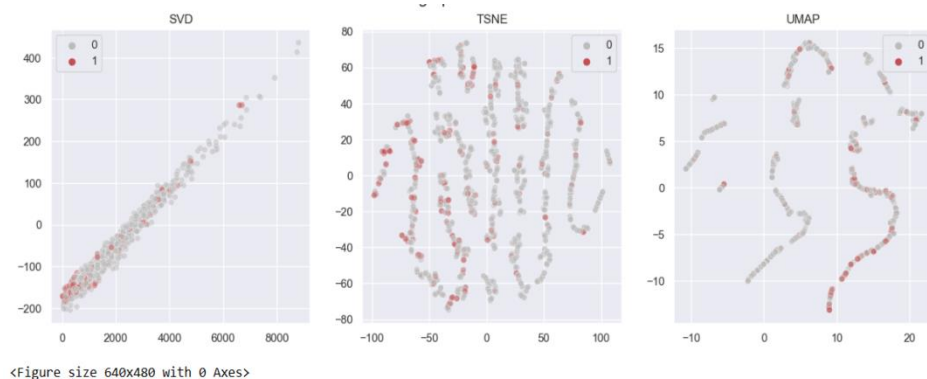


Figure 4: Vital Sign characteristics

- Laboratory features including hematocrit, red blood cells, Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), Mean Corpuscular Volume (MCV), Red blood Cell Distribution Width (RDW), platelet count, white blood cells, neutrophils, basophils, lymphocytes, Prothrombin Time (PT), and mean corpuscular hemoglobin concentration are clinically essential for screening and confirming the presence and level of disease risk in patients.

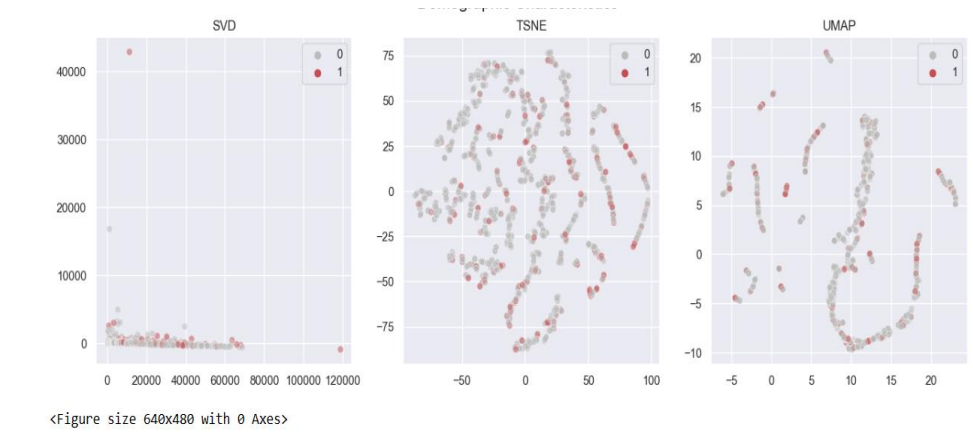


Figure 5: Laboratory Characteristics

3.3 Feature Engineering/ Feature Selection

Feature engineering involves manipulating and selecting relevant features from an unprocessed dataset to improve the performance of machine learning models. The initial tasks in the given code include addressing missing values by imputing the mean or mode for columns with missing data. The chosen characteristics include demographic information, comorbidities, vital signs, laboratory values, and the ejection fraction (EF) of the patient.

Following this, the dataset, denoted as *df*, is used in future modeling endeavors. The objective of this feature engineering procedure is to provide a more targeted and relevant set of features to predict in-hospital mortality among critical care unit patients with heart failure. Additionally, Recursive Feature Elimination (RFE) feature selection techniques were imported from the *sklearn.feature_selection* repository and implemented. RFE allows for the selection of the most important features that positively contribute to each model. The parameter 'n_features_to_select' is set to 30 out of 48 features in the dataset. The 'classifiers' dictionary contains the RFE results for each classifier, which executes RFE within that classifier. Each classifier starts RFE with a designated estimator and fits the scaled features.

In data analytics and machine learning workflows, the feature importance plot facilitates better insights and decision-making by providing a clear visualization of the most important features in the dataset. This technique is pivotal for interpreting model behavior and making informed decisions about feature contributions. To visualize feature importance, the study utilizes two Python libraries, *matplotlib* and *numpy*, due to their robust plotting capabilities. A horizontal bar plot showing the top 30 features and their importance is generated, sorted in descending order based on the significance of the features in the model, as shown in Figure 6.

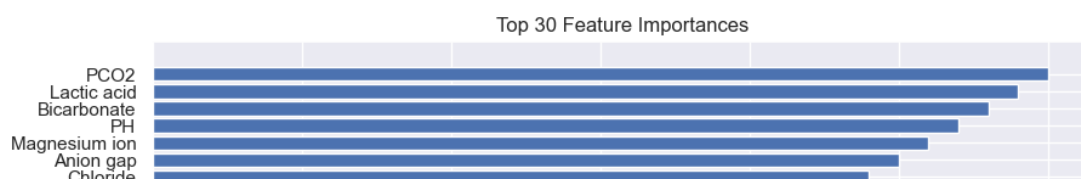


Figure 6: Feature Importance Plot

3.4 Model Training

In this research, model training involves the process of developing and evaluating predictive models for the risk of in-hospital mortality among heart failure patients in intensive care units (ICUs). Subsequently, the dataset is split into training (X_{train} , y_{train}) and testing (X_{test} , y_{test}) sets with the 'test_size= 0.2' (80% (941.6) for training the model and 20% (235.4) for testing the model), using the `train_test_split` function. To facilitate the identification of the best-performing model, the research considers a variety of classifiers, includes: RF, XGBoost, LGBost, and SVM. The `find_best_model` function iteratively trains each classifier on the training data and calculates its accuracy. The results are then compiled into a DataFrame, showcasing the accuracy scores of each model. This comprehensive approach enables the comparison of multiple algorithms, aiding in the selection of the most effective model for predicting in-hospital mortality among heart failure patients.

3.5 10-Fold Cross-validation

The k-fold cross-validation (CV) is a technique that assess the performance of the predictive model. The techniques involve partitioning the dataset into k equal-size subset, called k-folds. The model is then trained and evaluated k times, each time using a different fold as the validation set while the remaining folds are used for training.

The research considered 'k-fold_cross-validation' technique where in this case (k=10). Therefore, k-fold cross validation techniques in machine learning evaluates multiple algorithms simultaneously, in this case in 10-fold (10 splits) shuffling the data and settling a random state for reproducibility. Each iteration of cross-validation process computes 'cross_validation_score' for the model using accuracy as the scoring metrics. The process estimates how well these models perform on the unseen data (Omotihinwa *et al.*, 2023). This process typically helps in comparing the performance of different algorithms such as RF, SVM, LGBost and XGBoost. The process print the cross_validation scores and mean accuracy selecting the outperforming algorithm for the given task and further turning.

3.6 Model Testing (ROC and AUC)

Model testing in machine learning evaluates a trained model's generalization ability on fresh data. In this study, predictive models are trained on the training dataset and evaluated on the testing dataset (x test, y test) to determine their accuracy in predicting new occurrences. Model assessment is essential to confirm that the algorithm works effectively on fresh, real-world data and does not overfits training data. The study uses the Confusion Matrix and ROC Curves to evaluate performance. The Confusion Matrix shows the model's prediction performance, allowing computation of accuracy, precision, recall, and F1 score (Omotehinwa & Oyewola, 2023). ROC Curves show the trade-off between sensitivity and specificity, and the AUC-ROC measures the model's ability to distinguish positive and negative outcomes. These measurements demonstrate the model's ability to predict in-hospital mortality in critical care unit heart failure patients. The researchers used the confusion matrix to assess model accuracy.

Table 2: A confusion matrix for a binary classification task (AlBadani *et al.*, 2022)

Actual class	Predicted Class	
	Positive	Negative
True	True positive (TP)	False Negative (FN)
False	False positive (FP)	True Negative (TN)

The Confusion matrix describes the performance parameters for the classifier where TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively as shown in Table 2.

- **True Negative:** Observation is correctly classified as negative
- **False Negative:** observation is incorrectly classified as negative
- **True Positive:** A position is incorrectly classified by the model
- **False Positive:** A negative observation is incorrectly classified.

In calculating the accuracy, precision and recall for the model the following formula will be used:

i. **Accuracy measure:** Total number of correctly classified sentiments divided by total numbers of sentiments as shown in equation (1).

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

ii. **Precision:** Equation (2) shows the ability of the classifier to correctly identify the positive class.

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

iii. **Recall:** The number of times the classifier predicted a negative class out of all the times the class is negative (non-cancerous) as seen in equation (3).

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

F-Measure: It's the harmonic mean of the accuracy and recall. (4) defines F-Measure.

$$F - Measure = \frac{TP}{TP + \frac{1}{2} + FP + FN} \tag{4}$$

Receiver Operating Characteristics Curve (ROC): The ROC visualizes model performance. This comprehensive index measures continuous sensitivity and specificity.

Area under the curve (AUC): A model's performance is usually assessed using the area under the ROC curve, or AUC. AUC measures all two-dimensional ROC curve area. Model performance improves with greater AUCs.

4. Results and Discussion

4.1 Exploratory Data Analysis Results

Figure 7 below define the relationship between the feature and heart failure mortality best on the EDA results in demographic features age and BMI with less p-value of 0.026 and 0.067 are much significant and more related to the mortality in heart failure patients than the gender.

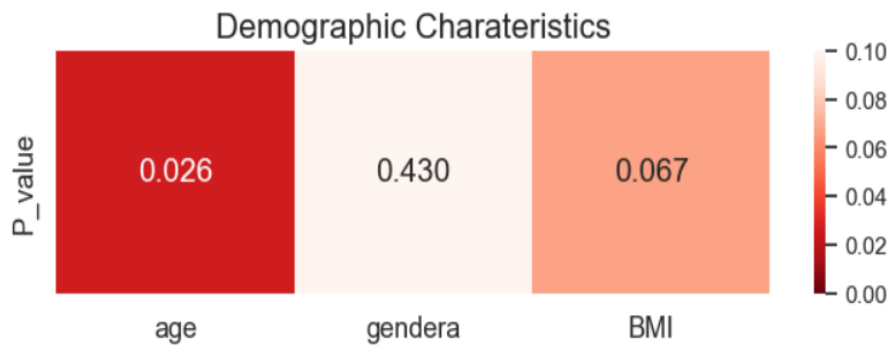


Figure 7: Demographic Characteristics

Also the EDA results in the comorbidities features (presence or absent of other diseases) atrial-fibrillation showed less p-value 0.000 among the other 9 features which recorded more than 0.5. The results indicates that only patients with atrial-fibrillation that need to be watch closely and more endangered to in-hospital mortality as shown in Figure 8 below.

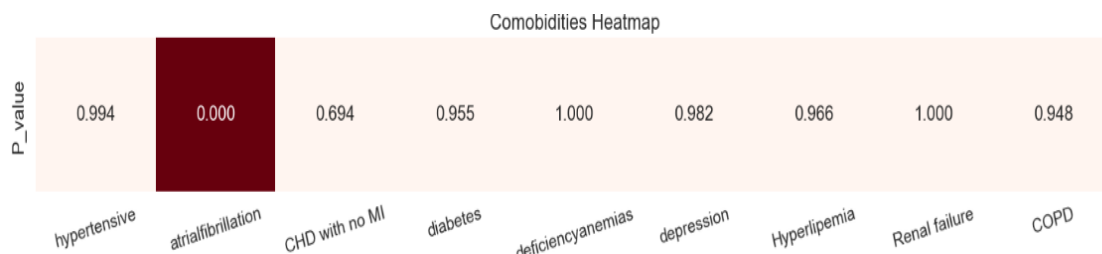


Figure 8: Comorbidities Characteristics

In vital sign features EDA results clearly indicates there was an amiable relationship between the heart rate, systolic blood pressure, diastolic blood pressure, respiratory rate, temperature, SP O2, and urine output with p-value of 0.000, 0.000, 0.003, 0.000, 0.002, 0.015, and 0.000 respectively with the risk of in-hospital mortality of heart failure patients as shown in figure 9 below.

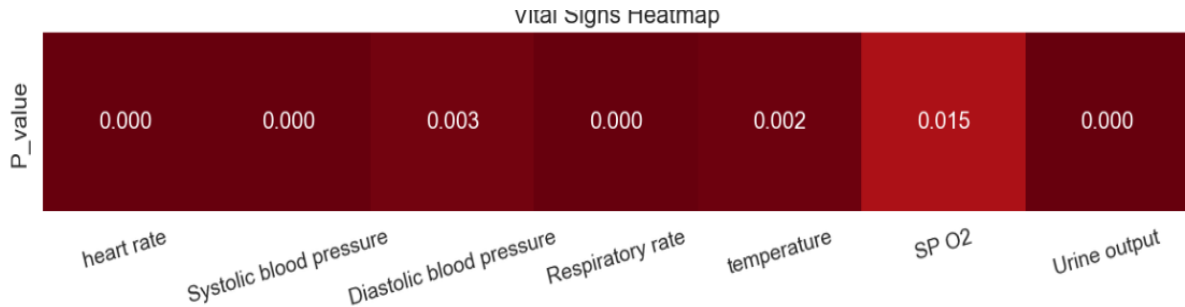


Figure 9: Vital Sign variables

The laboratory variable results of EDA showed RDW, Leucocyte, Platelets, Neutrophils, Basophils, Lymphocyte, PT, INR, NT-proBNP, Creatine kinase, Urea nitrogen, Blood potassium, Blood sodium, Blood calcium, Chloride, Anion gap, Magnesium ion, PH, Bicarbonate, Lactic acid, and PCO2 with the results recorded 0.000, 0.000, 0.002, 0.019, 0.000, 0.000, 0.000, 0.000, 0.015, 0.000, 0.000, 0.023, 0.000, 0.040, 0.000, 0.011, 0.000, 0.000, 0.000, and 0.000 respectively are significantly related to the risk of in-hospital mortality of heart failure patients as shown in Figure 10 below.

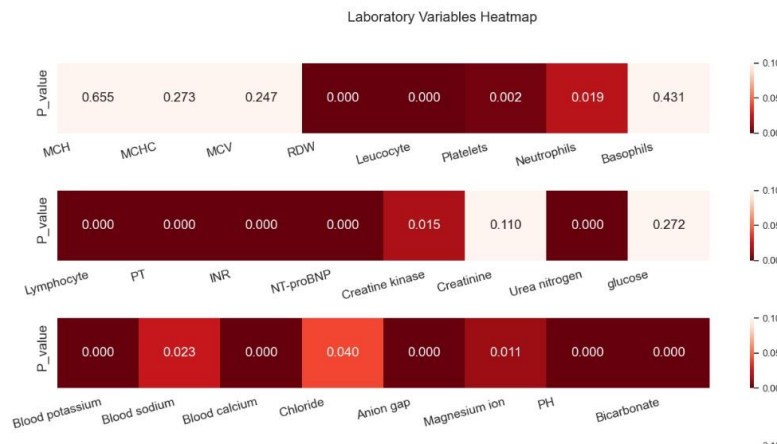


Figure 10: Laboratory variables

4.2 Model Evaluation Results:

The predictive analytics model for in-hospital mortality in heart failure patients within Intensive Care Units was developed and evaluated using Random Forest (RF), Support Vector Machine (SVM), LightGBM

(LGB), and XGBoost (XGB) algorithms, to accurately identify high-risk patients in crucial for improving outcomes and saving lives, to improve the medical outcome. The following results were obtained:

4.2.1 Model Performance Metric Result

Table 3: Model Performance Metric Result

Model	Accuracy	Precision	Recall	F1-score	Cross-Validation Accuracy	AUC-ROC	Confusion Matrix
Random Forest (RF)	0.94	0.95	0.93	0.94	0.94	0.99	[[202 10] [13 182]]
Support Vector Machine (SVM)	0.93	0.91	0.95	0.93	0.68	0.99	[[194 18] [9 186]]
LightGBM (LGB)	0.93	0.94	0.92	0.93	0.94	0.98	[[201 11] [16 179]]
XGBoost (XGB)	0.92	0.90	0.93	0.91	0.93	0.98	[[192 20] [14 181]]

Table 3 presents a comprehensive evaluation of different predictive analytics algorithms for assessing the risk of in-hospital mortality among heart failure patients in intensive care units. Four prominent algorithms, namely Random Forest (RF), Support Vector Machine (SVM), Light Gradient Boost (LGB), and Extreme Gradient Boost (XGB), were assessed based on key performance metrics, including Accuracy, Precision, Recall, and F1-score. Additionally, Cross-Validation Accuracy and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) were considered for a more thorough analysis. Among these models, the Support Vector Machine (SVM) and the Random Forest emerges as the best-performing algorithms, showcasing the highest Precision (0.96), Recall (0.93) and an impressive F1-score (0.90). Moreover, RF exhibits robust generalization with a Cross-Validation Accuracy of 0.94 and an outstanding AUC-ROC value of 0.99. These results highlight RF's effectiveness in predicting the risk of in-hospital mortality in intensive care units for heart failure patients, making it the preferred algorithm for this critical healthcare application.

4.2.2 AUC & ROC plot:

The AUROC curves Random Forest and the Support Vector Machine showed the best performance with ninety-nine percent (99%) accuracy compared to LGB and XGB with ninety-eight percent (98%) accuracy.

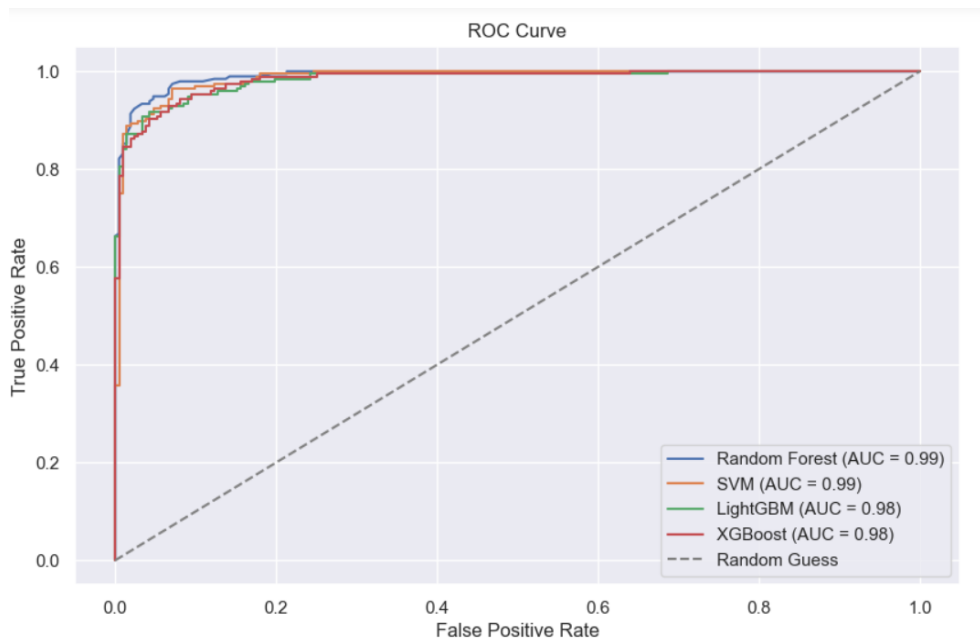


Figure 11: AUC & ROC results

5. Conclusion

This research work is aimed at determining a more efficient prediction model between RF, SVM, LGB and XGB to predict the risk of in-hospital mortality of heart failure patients in the intensive care unit. Using the models, can help in understanding the symptoms that are highly risk and contribute to in-hospital mortality of heart failure patients. The study's primary goal is to develop a suitable and accurate predictive model that can be used by researchers and clinical workers to predict and manage the patients with high risk of in-hospital mortality in the ICU. Also the exploratory data analysis carried out in this study will help clinicians identify the potential clinical symptoms that are highly related to heart failure patients and their risk of in-hospital mortality. Despite the constraints, the study came up with one model among the four (Random Forest) that exhibits exceptional performance with perfect results in both the 10-fold cross-validation estimation techniques and classification metrics such as: accuracy, precision, and recall, highlighting its unique classification and prediction proficiencies.

5.1 Recommendation

From the foregoing discussions, the study and particularly the model is recommended for use by researchers and healthcare workers who are into the prediction of the risk of in-hospital mortality especially in the critical unit of ICU to help in addressing the problems related to high risk of in-hospital mortality rate in patients heart failure complications within the ICU's. Additionally, it can provide a synergistic approach that improves prediction accuracy, speeds up the process, and reduce the potential dangers associated with poor predictions.

5.2 Future Work

As the scope of this work is limited to retrospective cohort data from MIMIC-III, using four predictive models (RF, SVM, LGB, and XGB) and evaluates with classification metrics of Accuracy, Precision, Recall and F1_score, Cross Validation estimation techniques and also conduct additional evaluation metrics such as confusion matrices, Receiver Operating Characteristic (ROC) Curves, or Area Under the Curve (AUC)

to gain a more comprehensive understanding of the models' performance. Gather a more diverse and representative dataset to enhance the training and generalization capabilities of the models.

Explore other machine learning algorithms or ensemble methods to compare their performance with the existing models, considering factors such as interpretability and computational efficiency.

Reference

- Almerund, S., Alapack, R. J., Fridlund, B., & Ekebergh, M. (2007). Vigilance and Invisibility-Being a Patient in Technology Intense Environments. *Nursing in Critical Care*, 12(3), 151-158.
- Akpan, V. A., Omotehinwa, T. O., & Agbogun, J. B. (2022). Adaptive Classification of Impending Human Heart, Kidney and Liver Failures Based on Measurable Blood-Related Parameters Using MIMO HANFA-ART with ACA Algorithms. *Biomedical Sciences*, 8(3), 97–112. <https://doi.org/10.11648/J.BS.20220803.14>
- Bozkurt, B., Coats, A., Tsutsui, H., Abdelhamid, C., Adamopoulos, S., Albert, N., Anker, S., Atherton, J., Böhm, M., Butler, J., Drazner, M., Felker, G., Filippatos, G., Fiuzat, M., Fonarow, G., Gómez-Mesa, J., Heidenreich, P., Imamura, T., Jankowska, E. (2021). Universal definition and classification of heart failure: a report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society. *European Journal of heart failure*, 23. doi:<https://doi.org/10.1002/ejhf.2115>.
- Chang PP, Wruk LM, Shahar E, . (2018). Trend in Hospitalizations and Survival of Acute Decompensated Heart Failure in four US Communities. *Reseach Gate*.
- Cheng, Z., Li, T., Zeng, D., & Wang, K. (2023). Mechine Learning-Based In-hospital Mortality Risk Prediction tool for Intensive Care Unit Patients with Heart Failure. *Frontiers in Cardiovascular Medicine*. doi:10.1119699.
- Dineo M., Turgay C., Eric K., & Hopewell N. (2022). Predicting in-hospital all-cause mortality in heart failure using machine learning. *PMC PubMed Central*, 9:1032524. doi:10.3389/fcvm.2022.1032524
- Fen M., Yun-Peng C., Yu-Xiao Z., Xiao-Mao F., & Ye L. (2018). Predictive Modeling of Hospital Mortality for Patients with Heart. *IEEE Access*. doi:10.1109/ACCESS.2018.2789898.
- Guilan K., Ke L., & Younhua H. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *Research Gate*, 20(1:251. doi:<https://doi.org/10.1186/s12911-020-01271-2>
- Hsich, E., Gorodeski, E., Blackstone, E., Ishwaran, H., & Lauer, M. (2011). Identifying Important Risk Factors for Survival in Patient With Systolic Heart Failure Using Random Survival Forests. *Circulation: Cardiovascular Quality and Outcomes*. doi:<https://doi.org/10.1161/CIRCOUTCOMES.110.939371>.
- Jacob A. U., Hong L., & Donald A. R. (2017). Failure of fertility therapy and subsequent adverse cardiovascular events. *PubMed Central*, 189(10):E391-E397. doi:<https://doi.org/10.1503/cmaj.160744>.
- Jili L., Siru L., Yundi H., Lingfeng Z. Yujja M. & Jianlin L. (2022). Predicting Mortality in Intensive Care Unit Patients With Heart Failure Using an Interpretable Machine Learning Model: Retrospective Cohort Study. *Journal of Medical Internet Research*, 24(8): e38082. doi:10.2196/38082
- Jordan, M., & Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. *Journal of Science*, 349, 255 - 260. doi:<https://doi.org/10.1126/science.aaa8415>.

- Khera, R., Haimovich, J., Hurley, N. C., McNamara, R., Spertus, J. A., Desai, N., & Krumholz, H. M. (2021). Use of Machine Learning Models to Predict death after Acute Myocardial infarction. *JAMA Cardiology*, 6(6), 633-641.
- Lee YJ, Cho KJ, Kwon O, Park H, Lee Y, Kwon JM, Park J, Kim JS, Lee MJ, Kim AJ, Ko RE, Jeon K, Jo YH. A. (2021). multicentre validation study of the deep learning-based early warning score for predicting in-hospital cardiac arrest in patients admitted to general wards. *Resuscitation*, 163:78-85. doi:10.1016/j.resuscitation.2021.04.013. Epub ahead of print. PMID: 33895236.
- Li, F., Xin, H., Zhang J, Fu, M., Zhou, J., Lian, Z. (2021). Prediction Model of In-hospital Mortality in Intensive Care Unit Patients with Heart Failure: Machine learning based, Retrospective analysis of the MIMIC-III Database. *BMJ Open*, 11(7). doi:9. doi:10.1136/
- Mathew W. S., Muthiah V., Kershaw P., Darren K. M., Javed B., Gregg F., Mujjeb B., Vaishnavi K., Justin L. G., Brendan E., Duwayne W., Jarett B., & Ambarish P. (2019). Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score. *PMC PubMed Central*, 42(12): 2298–2306. . doi:10.2337/dc19-0587.
- Mohan L, R. N. (2023). Hydrocortisone-associated death and hospital length of stay in patients with sepsis: A retrospective cohort of large-scale clinical care data. *Science Direct* .
- Mohd J, Abid H., Ravi P., Suman R., Shanay R. (2022). Significance of Machine Learning in Healthcare: Features, Pillars and applications. *Science Direct*, 58-73. doi:<https://doi.org/10.1016/j.ijin.2022.05.002>.
- Naghme K., Zohreh K. & Mohammad J. T. (2021). The prediction of mortality influential variables in an intensive care unit: a case study. *SpringerLink*, 203–219. doi:<https://doi.org/10.1007/s00779-021-01540-5>
- Omotehinwa, T. O., Oyewola, D. O., & Dada, E. G. (2023). A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis. *Healthcare Analytics*, 4, 100218. <https://doi.org/10.1016/J.HEALTH.2023.100218>
- Omotehinwa, T. O., & Oyewola, D. O. (2023). Hyperparameter Optimization of Ensemble Models for Spam Email Detection. *Applied Sciences (Switzerland)*, 13(3), 1971. <https://doi.org/10.3390/AP13031971>
- Patrick E., Baran B., & Vignesh S.,. (2020). Decompensation in Critical Care: Early Prediction of Acute Heart Failure Onset. *JMIR Publications*. doi:10.2196/19892
- Ponikowski P, Voors AA, Anker SD, . (2016). The Guidelines for the diagnosis and treatment of acute and chronic heart failure:. *The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology*, (p. 200). EU.
- Rabi M., Yazid AH, Muhammad A.A., Habeebah AK. (2024). Heart Failure Prediction Using Variational Auto-Encoder and Extreme Gradient Boosting Neural Network. *KASU Journal of Computer Science (KJCS)*. Retrieved from https://www.kjcs.com.ng/view_paper/19
- Raphael K., Carl O., Martin W., Thomas T., Hedi M., Helle A. & Tim B. (2022). Towards definitions of critical illness and critical care using concept analysis. *BMJ Open*, 12(9): e060972.
- Smith, R. H. (n.d.). *University of Maryland*. Retrieved from Robert Smith School of Business Analytics: <https://onlinebusiness.umd.edu/blog/what-is-predictive-analytics-and-why-is-it-important/>
- Tetana, T. (2016). *Assessing the risk of Coronary heart disease among rural adults*. Eastern Cape: Joe Agabi.

- Tian P, Liang L, Zhao X, Huang B, Feng J, Huang L, Huang Y, Zhai M, Zhou Q, Zhang J, & Zhang Y. (2023). Machine Learning for Mortality Prediction in Patients With Heart Failure With Mildly Reduced Ejection Fraction. *J Am Heart Assoc.*, 20;12(12). doi:e029124. doi::https://doi.org/ 10.1161/JAHA.122.029124.
- Veli-Pekka H., Alexandre M., Jelena Č., Dominique B., Hector B., Ovidiu C., Maria G. C., Volkmar F., Gerasimos F., Simon G., Adelino Leite-M., Johan L., Josep M., Christian M., Wilfried M., Robert N., Anton V. N., John P., Jillian P. R., Arsen R., & Gius. (2016). Contemporary management of acute right ventricular failure: a statement from the Heart Failure Association and the Working Group on Pulmonary Circulation and Right Ventricular Function of the European Society of Cardiology. *European Journal of Heart Failure*, 18(3):226-41. doi::https://doi.org/10.1002/ejhf.478.
- Weintraub, W., Fahed, A., & Rumsfeld, J. (2018). Translational Medicine in the Era of Big Data and Machine Learning. *Circulation Research*. . doi:https://doi.org/10.1161/CIRCRESAHA.118.313944.
- Yang L., Zhiyu W., & Cong W. (2021). Improvement of APACHE II score system for disease severity based on XGBoost algorithm. *BMC Med Inform Decis Mak*, 21: 237. doi:10.1186/s12911-021-01591-x.