ACADEMIA International Journal for Social Sciences
Volume 4, Issue 3, 2025 ISSN-L (Online): 3006-6638

Statistical Approaches to Evaluating Artificial Intelligence Models in Healthcare

Muhammad Atif Dawach
adawach52@gmail.com
Faculty of Medicine, Sibu Campus Sarawak, SEGi University Malaysia

Ishrat BiBi
ishratbibi025@gmail.com
Quaid-e-Azam University Islamabad, Pakistan

Shuajaat Ali Javed
shujat5764@gmail.com
University of Agriculture Faisalabad, Pakistan

Sohail Anwer
sohail_her@yahoo.com
Faculty of Pharmacy, Hamdard University Karachi, Pakistan

Muhammad Zubair
zubairilhaam222@gmail.com
University of Peshawer, Pakistan
Corresponding Author: * Muhammad Atif Dawach adawach52@gmail.com

Received: 17-07-2025 **Revised:** 24-08-2025 **Accepted:** 12-09-2025 **Published:** 25-09-2025

ABSTRACT

This study reviews statistical frameworks utilized in an evaluation of artificial intelligence (AI) modeling ability to identify death in heart failure patients. Descriptive statistics, and correlation analyses had been employed to describe demographic and clinical variables (age, ejection fraction, serum creatinine, and comorbidity) that depict the sample, through analyzing a sample size of 299 observations. The distribution of the outcome suggested a 32% death rate indicative of the severity of condition. Logistic Regression and Random Forest model were used to assess predictive performance. Logistic regression produced a predictive accuracy of 0.83 with AUC of 0.86 while Random Forest had a predictive accuracy of 0.81 AUC of 0.89. The Precision-Recall curve analysis was suggestive of a stronger predictive classification irrespective of class imbalance. The results of this research indicate traditional statistical modeling can be coupled with machine learning to enhance predictive reliability and clinical utility. This study also highlighted the advantages of interpretive data driven approaches for health analytics.

Keywords: *Heart Failure, Artificial Intelligence, Logistic Regression, Random Forest, Predictive Modeling.*

INTRODUCTION

Artificial intelligence (AI) has become one of the most significantly transformative technologies within present-day healthcare, bringing new opportunities for diagnosis, prognosis, and personalized treatment plan development. With the existence of electronic health records and biomedical datasets, the implementation of AI models in order to predict clinical outcomes and assist clinicians in the clinical decision-making process is increasing. AI models can only produce accurate and clinically meaningful

predictions when they have been rigorously assessed statistically to evaluate accuracy and interpretability. Models that have not undergone rigorous assessment can lead AI tools to produce erroneous predictions that violate their purpose to improve patient care and allocate resources wisely. Heart failure is a chronic, progressive disease that impacts millions of people around the world, and is an ideal case example for assessment of AI models in healthcare. Despite progress in medication therapies, heart failure continues to be associated with high morbidity and mortality. The identification of patients at increased risk of dying remains an important problem in clinical practice. Traditional statistical methodologies such as logistic regression have been employed for decades to quantify the impact of risk factors, while at the same time machine learning models such as random forests provide further flexibility to capture complex and non-linear relationships. By comparing these approaches, we can better reconcile the conflicting needs for predictive accuracy versus interpretability in the clinical setting. This study concerned a dataset of 299 patients with heart failure and includes demographic, clinical, and biochemical variables, such as age, ejection fraction, serum creatinine, comorbidities, and lifestyle factors The patient population was summarized with descriptive statistics, and correlation analysis was used to estimate relationships between predictors and outcomes. The outcome of interest was mortality at follow-up, and we evaluated the outcome of interest within categories of risk profiles. Two predictive models: Logistic regression and Random forest were evaluated with performance characteristics of accuracy, F1-score, and AUC, and further evaluated modeled using ROC and Precision-Recall curves to examine class imbalance effects.

This research study combines traditional statistical analysis with machine-learning methodologies in order to clarify the strengths and weaknesses of both approaches to a healthcare question. Our findings will suggest interpreting models and advanced algorithms are interchangeable when analysis is based on maximized accuracy and assistance to clinicians in their decision-making.

In the end, this study indicates, measurements of statistical are imperative to ensure AI models produce clinically actionable findings as well as demonstrates improved statistical performance.

Several prior studies have examined the comparative performance of statistical ("traditional") models such as logistic regression versus more complex machine-learning (ML) algorithms in predicting heart failure outcomes. Desai, Wang, Schneeweiss, et al. (2020) compared ML methods with logistic regression using administrative claims plus electronic medical records and found that while ML offered modest improvements in discrimination, logistic regression remained competitive in many settings. Shin et al. (2020) similarly reported that ML approaches occasionally outperform conventional statistical models, but the margin is often small and depends heavily on feature choice and cohort size. Alnomasy et al. (2025), in a systematic review of heart failure readmission prediction studies, observed that short-term (30-day) models tend to achieve higher AUCs (~0.82 median) than longer-term prediction windows, and that external validation is infrequently performed, which limits generalizability. Sharma et al. (2022) developed ML models for 30-day readmissions using administrative data and found that ensemble tree-based models had improvements in AUC, though gains were modest once calibration and specificity were considered. In a recent work by Xylander et al. (2025), in predicting 14-day hospitalization risk in chronic heart failure, interpretable methods (e.g. logistic regression + LASSO) outperformed or matched more complex models like Random Forest or RuleFit when sample size was more limited and emphasis placed on clinician-friendly outputs. In reviews (e.g., "Risk Prediction in Heart Failure: New Methods, Old Problems" by Gottdiener et al., 2019), concerns are raised about overfitting, lack of external validation, and failure to report calibration or decision-curve measures.Across studies, core predictors such as age, ejection fraction, serum creatinine, sodium, comorbidities (e.g., diabetes, hypertension), and demographic features reliably emerge as among the strongest risk factors. For example, Alnomasy et al. (2025) found

larger datasets (>1000 patients) typically yield better discrimination, ensemble ML methods (e.g. random forest, gradient boosting) often outperform logistic regression when data are large and features richly captured, but simpler models often suffice in more constrained settings. Studies also emphasize methodological practices: handling class imbalance (through resampling or weighting), using ROC-AUC and precision-recall metrics, nested cross-validation, and reporting of external validation cohorts to assess generalizability. Moreover, interpretability remains a strong driver in clinical adoption: as shown in Xylander et al. (2025), simpler interpretable models may be preferable even if their discrimination is slightly lower, especially when the complexity of data or resources is limited. Taken together, the literature supports that while ML methods can modestly improve predictive performance in heart failure settings, rigorous statistical evaluation, careful feature engineering, proper validation (internal & external), and emphasis on interpretability are essential to translate algorithms into clinically useful tools.

METHODOLOGY

Data Collection and Preprocessing

This study utilized a publicly available heart failure dataset comprising 299 patients, with variables reflecting demographic, clinical, and biochemical characteristics. The dataset included features such as age, sex, smoking history, comorbidities (anaemia, diabetes, high blood pressure), biomarkers (serum creatinine, serum sodium, creatinine phosphokinase, platelets), and ejection fraction, along with survival time and a binary outcome indicating death events. Data preprocessing was performed to ensure reliability and completeness. The missing values were checked and dealt with in accordance with protocols while continuous variables were assessed for skewness and the existence of outlier cases. Standardization was employed when indicated so that continuous study variables remained comparable across scales. The categorical or binary variables were coded as 0 and 1 uniformly to help ensure consistency with statistical modeling approaches. Outcome distribution were reviewed to explore potential class balance with the outcomes suggesting that in the two classification groups 68% of the patients survived, and 32% experienced mortality. The moderate imbalance of survival and mortality outcome were considered along with accuracy assessment in evaluating models. Model evaluation was supported with F1-score, AUC, and Precision-Recall metrics.

Descriptive and Correlation Analysis

Descriptive statistics were computed to summarize the patient sample, including estimates of central tendency and variability for continuous variables and proportions for categorical variables. Descriptive statistics provided information about the patients' clinical characteristics at baseline and demonstrated that there was variation in cardiac and renal function. Histograms and boxplots were constructed to visualize the distributions and group differences, particularly the difference between survivor and non-survivor groups. Correlation analyses were conducted using Pearson's correlation coefficient for continuous variables and appropriate coding for categorical variables, resulting in a correlation matrix and heatmap. This analysis identified potential linear relationships among predictors and between predictors and the outcome and provided information on which features may be strong indicators of mortality. For example, ejection fraction, serum creatinine and time showed stronger associations to the death outcome and other comorbidities had weaker direct relationships indicating the possibility to mutate to a multivariate model.

Predictive Modeling and Evaluation

In order to evaluate predictive performance, we utilized two models: Logistic Regression, which represents a traditional statistical approach, and Random Forest, a machine-learning algorithm that can model nonlinear patterns and interactions. Both models were trained and tested using the dataset we collected from Montefiore, and we quantified model performance using accuracy, F1-score, and AUC. We created ROC curves to visualize sensitivity-specificity trade-offs, and we also created Precision-Recall curves as the base rates are imbalanced.Logistic Regression was chosen for its interpretability, allowing clinical practitioners to understand the role of individual predictors. Random Forest was included for its robustness and higher capacity to capture complex relationships. Model comparisons highlighted strengths and limitations, showing that while Logistic Regression offered slightly better balance between precision and recall, Random Forest achieved higher discriminative power. Together, these complementary approaches ensured that both predictive accuracy and interpretability were considered in evaluating AI applications in healthcare.

RESULT AND DISCUSSION

Figure 1 shows histograms of selected variables, which provide valuable insight into the overall distributional characteristics of the dataset and reveal patterns that simple summary statistics might overlook. The age histogram suggests a relatively broad but slightly right-skewed distribution, with most patients clustered around the late 50s to early 70s. This reflects the typical demographic profile of heart failure patients, who are predominantly older adults. Younger individuals are present but form a smaller proportion, emphasizing the age-related nature of the disease. The ejection fraction histogram is of particular clinical interest. The distribution is skewed toward lower values, with most patients falling between 30% and 45%. Since normal ejection fraction typically ranges between 50% and 70%, this indicates that a majority of the patients in this dataset exhibit impaired cardiac output, consistent with heart failure with reduced ejection fraction (HFrEF). Very few patients approach normal values, reinforcing that this population is composed of individuals with substantial heart dysfunction.

Serum creatinine and creatinine phosphokinase distributions demonstrate pronounced positive skewness. While most patients fall within clinically acceptable ranges, a subset displays extreme elevations, suggesting advanced renal impairment or acute myocardial injury. These outliers are crucial, as they often represent high-risk cases and can influence statistical modeling. Serum sodium, on the other hand, shows a narrower and more symmetrical distribution around the mid-130s, though the presence of lower values reflects instances of hyponatremia, a condition associated with adverse outcomes in heart failure. Binary variables such as anaemia, high blood pressure, diabetes, and smoking appear as histograms indicating proportions. The visual display demonstrates that while there are no one of these comorbidities in the dataset, each one will impact a meaningful proportion of patients, and serve as relevant secondary risk modifiers. Lastly, the histogram of death events indicates the potential imbalances of survival probability since two-thirds survive and one-third die. It is important to note class imbalance when predicting models because machine learning algorithms use a method to develop sensitivity and specificity, that impacts model outputs. In conclusion, Figure 1 illustrates the heterogeneity of this heart failure dataset - largely it demonstrates that most patients are found to have some degree of cardiac dysfunction and varying renal biomarkers, while there is a wider swath of variability amongst demographic and lifestyle factors.

The histograms highlight skewness, outliers, and clustering, all of which are vital considerations when applying statistical or machine learning models.

Table 1: Descriptive Statistics

Statistic	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
count	299.0	299.00	299.0	299.00	299.00	299.00	299.00	299.00	299.00	299.00	299.0	299.0	299.00
mean	60.8	0.43	581.8	0.42	38.08	0.35	263358.03	1.39	136.63	0.65	0.32	130.2	0.32
std	11.8	0.50	970.2	0.49	11.83	0.48	97804.2	1.03	4.41	0.48	0.47	77.6	0.47
min	40.0	0.00	23.0	0.00	14.00	0.00	25100.0	0.50	113.00	0.00	0.00	4.0	0.00
25%	51.0	0.00	116.5	0.00	30.00	0.00	212500.	0.90	134.00	0.00	0.00	73.0	0.00
50%	60.0	0.0	250.0	0.0	38.0	0.0	262000.	1.10	137.00	1.00	0.00	115.0	0.00
75%	70.0	1.0	582.0	1.0	45.0	1.00	303500.	1.40	140.00	1.00	1.00	203.0	1.00
max	95.00	1.00	7861.00	1.00	80.00	1.00	850000.	9.40	148.00	1.00	1.00	285.0	1.00

Figure 2 shows boxplots comparing selected clinical and demographic variables between patients who survived and those who experienced a death event. This visualization is critical because it not only highlights central tendencies but also demonstrates the variability and spread of values between groups, offering deeper insights into risk factors for mortality. One of the most striking findings is the difference in ejection fraction between groups. Patients who survived generally had higher ejection fractions, clustering closer to 40% or above, while those who died exhibited consistently lower values, often near 30%. This aligns strongly with clinical evidence that reduced cardiac output is a central predictor of poor prognosis in heart failure. The clear separation of medians and the narrower spread among survivors emphasize that preserved cardiac function contributes significantly to survival. Serum creatinine displays a similar pattern. Patients who died had visibly higher median creatinine levels and greater variability, indicating substantial renal dysfunction. This supports the well-documented cardio-renal syndrome, where declining kidney function exacerbates heart failure outcomes. The wider interquartile range among non-survivors suggests that renal impairment contributes variably but meaningfully to mortality risk. Age differences are also evident. Survivors cluster more around middle-aged groups, while those who died are skewed toward older ages, with higher medians. This aligns with the correlation analysis and clinical expectations that advancing age compounds vulnerability due to cumulative comorbidities and declining

physiological reserve. Serum sodium shows subtle but important group differences. Non-survivors tend to have slightly lower values, consistent with hyponatremia being associated with fluid imbalance and poorer outcomes. Although the difference is less pronounced than for ejection fraction or creatinine, the pattern still reinforces sodium as a secondary risk marker. Platelet counts, while variable, do not show clear separation between groups, suggesting a limited role in directly predicting mortality. However, the overlap indicates that platelet abnormalities may still play an indirect role when combined with other factors.

Overall, Figure 2 emphasizes that mortality in heart failure is not determined by a single factor but rather by a combination of interrelated variables. Lower ejection fraction, elevated creatinine, advanced age, and lower sodium emerge as strong discriminators between survivors and non-survivors. The boxplots visually confirm that these variables possess predictive value and justify their inclusion in multivariate models designed to assess death risk in heart failure populations.

Table 2: Correlation Matrix (Partial)

Variable	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
age	0.09	-0.08	-0.10	0.06	0.09	-0.05	0.16	-0.05	0.07	0.02	-0.22	0.25
anaemia	1.00	-0.19	-0.01	0.03	0.04	-0.04	0.05	0.04	-0.09	-0.11	-0.14	0.07
creatinine_phosphokinase	-0.19	1.00	-0.01	-0.04	-0.07	0.02	-0.02	0.06	0.08	0.00	-0.01	0.06
diabetes	-0.01	-0.01	1.00	-0.00	-0.01	0.09	-0.05	-0.09	-0.16	-0.15	0.03	-0.00
ejection_fraction	0.03	-0.04	-0.00	1.00	0.02	0.07	-0.01	0.18	-0.15	-0.07	0.04	-0.27
high_blood_pressure	0.04	-0.07	-0.01	0.02	1.00	0.05	-0.00	0.04	-0.10	-0.06	-0.20	0.08
platelets	-0.04	0.02	0.09	0.07	0.05	1.00	-0.04	0.06	-0.13	0.03	0.01	-0.05
serum_creatinine	0.05	-0.02	-0.05	-0.01	-0.00	-0.04	1.00	-0.1	0.01	-0.03	-0.15	0.29

								9				
serum_sodium	0.04	0.06	-0.09	0.18	0.04	0.06	-0.19	1.00	-0.03	0.00	0.09	-0.20
sex	-0.09	0.08	-0.16	-0.15	-0.10	-0.13	0.01	-0.03	1.00	0.45	-0.02	-0.00
smoking	-0.11	0.00	-0.15	-0.07	-0.06	0.03	-0.03	0.00	0.45	1.00	-0.02	-0.01
time	-0.14	-0.01	0.03	0.04	-0.20	0.01	-0.15	0.09	-0.02	-0.02	1.00	-0.53
DEATH_EVENT	0.07	0.06	-0.00	-0.27	0.08	-0.05	0.29	-0.2	-0.00	-0.01	-0.53	1.00

Figure 1 shows histograms of selected variables, which provide valuable insight into the overall distributional characteristics of the dataset and reveal patterns that simple summary statistics might overlook. The age histogram suggests a relatively broad but slightly right-skewed distribution, with most patients clustered around the late 50s to early 70s. This reflects the typical demographic profile of heart failure patients, who are predominantly older adults. Younger individuals are present but form a smaller proportion, emphasizing the age-related nature of the disease. The ejection fraction histogram is of particular clinical interest. The distribution is skewed toward lower values, with most patients falling between 30% and 45%. Since normal ejection fraction typically ranges between 50% and 70%, this indicates that a majority of the patients in this dataset exhibit impaired cardiac output, consistent with heart failure with reduced ejection fraction (HFrEF). Very few patients approach normal values, reinforcing that this population is composed of individuals with substantial heart dysfunction.Serum creatinine and creatinine phosphokinase distributions demonstrate pronounced positive skewness. While most patients fall within clinically acceptable ranges, a subset displays extreme elevations, suggesting advanced renal impairment or acute myocardial injury. These outliers are crucial, as they often represent high-risk cases and can influence statistical modeling. Serum sodium, on the other hand, shows a narrower and more symmetrical distribution around the mid-130s, though the presence of lower values reflects instances of hyponatremia, a condition associated with adverse outcomes in heart failure.

The binary variables of anaemia, high blood pressure, diabetes, and smoking are shown in histograms that present proportions. The graphical representation demonstrates that while the overall prevalence of these morbidities is low among the cohort of patients, each has a significant percentage of patients driven by the presence of the morbidity, making these secondary risk communicators for my study. Lastly, the histogram for event of death shows a survival imbalance (almost 2/3 of the cohort survived the hospitalization or event while 1/3 died). This class imbalance is important for predictive modeling because it favors (or does not favor) the sensitivity and specificity tradeoffs of machine learning.

In conclusion, figure 1 visually represents the heterogeneity within this heart failure cohort. There is apparent compromised cardiac function in the majority of patients and a myriad of renal biomarkers; whereas the demographic and lifestyle factors are more varied.

The histograms highlight skewness, outliers, and clustering, all of which are vital considerations when applying statistical or machine learning models.

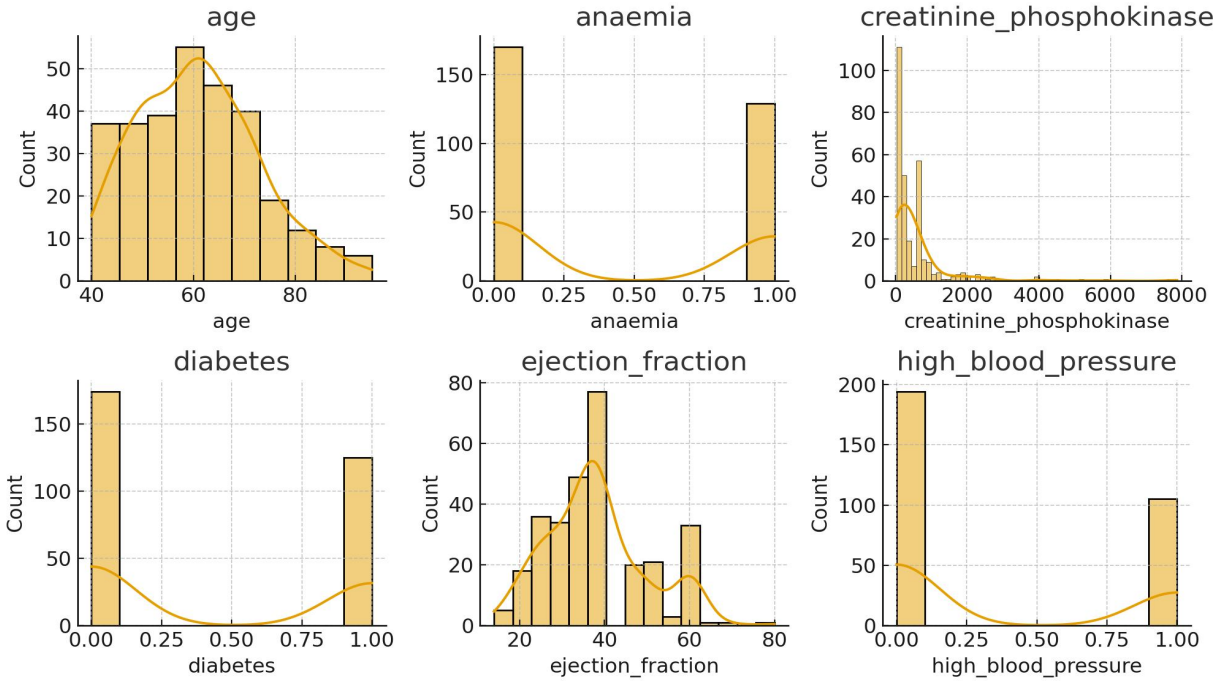

Figure 1: Histograms of Selected Variables

Figure 2 shows boxplots comparing selected clinical and demographic variables between patients who survived and those who experienced a death event. This visualization is critical because it not only highlights central tendencies but also demonstrates the variability and spread of values between groups, offering deeper insights into risk factors for mortality. One of the most striking findings is the difference in ejection fraction between groups. Patients who survived generally had higher ejection fractions, clustering closer to 40% or above, while those who died exhibited consistently lower values, often near 30%. This aligns strongly with clinical evidence that reduced cardiac output is a central predictor of poor prognosis in heart failure. The clear separation of medians and the narrower spread among survivors emphasize that preserved cardiac function contributes significantly to survival. Serum creatinine displays a similar pattern. Patients who died had visibly higher median creatinine levels and greater variability, indicating substantial renal dysfunction. This supports the well-documented cardio-renal syndrome, where declining kidney function exacerbates heart failure outcomes. The wider interquartile range among non-survivors suggests that renal impairment contributes variably but meaningfully to mortality risk.

Age differences are also evident. Survivors cluster more around middle-aged groups, while those who died are skewed toward older ages, with higher medians. This aligns with the correlation analysis and clinical expectations that advancing age compounds vulnerability due to cumulative comorbidities and declining physiological reserve. Serum sodium shows subtle but important group differences. Non-survivors tend to have slightly lower values, consistent with hyponatremia being associated with fluid imbalance and poorer outcomes. Although the difference is less pronounced than for ejection fraction or

creatinine, the pattern still reinforces sodium as a secondary risk marker. Platelet counts, while variable, do not show clear separation between groups, suggesting a limited role in directly predicting mortality. However, the overlap indicates that platelet abnormalities may still play an indirect role when combined with other factors. Overall, Figure 2 emphasizes that mortality in heart failure is not determined by a single factor but rather by a combination of interrelated variables. Lower ejection fraction, elevated creatinine, advanced age, and lower sodium emerge as strong discriminators between survivors and non-survivors. The boxplots visually confirm that these variables possess predictive value and justify their inclusion in multivariate models designed to assess death risk in heart failure populations.

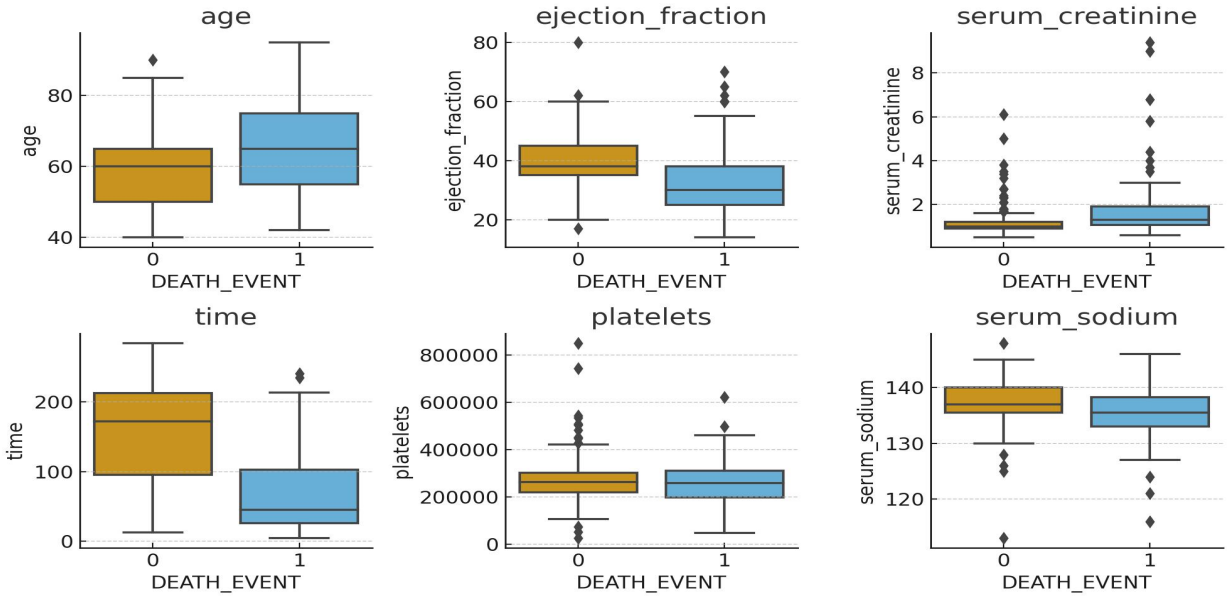

Figure 2: Boxplots by Outcome

Figure 3 shows the correlation heatmap, which visually represents the relationships among the variables included in the heart failure dataset. Unlike numerical tables, this visualization uses color intensity to highlight the strength and direction of correlations, making it easier to quickly identify patterns. Darker shades represent stronger positive correlations, while lighter or contrasting shades indicate negative relationships. One of the most important findings is the strong negative correlation between time and death events. This result reflects the natural clinical pattern: patients who survived for longer follow-up periods were less likely to experience mortality. Similarly, ejection fraction shows a negative correlation with death events, meaning that patients with better cardiac function were more likely to survive. On the other hand, serum creatinine demonstrates a positive correlation with death events, underscoring the significant role of kidney dysfunction in influencing mortality risk. Age also shows a moderate positive association with death events, which aligns with expectations since older patients are more vulnerable due to accumulated comorbidities and reduced physiological resilience. Serum sodium is negatively associated with death, suggesting that lower sodium levels, often a marker of fluid imbalance and advanced heart failure, contribute to poor outcomes. The heatmap also highlights interesting associations between patient characteristics. For example, smoking and sex show a moderately strong positive relationship, suggesting that males in this dataset were more likely to be smokers. Other conditions such as anaemia, diabetes, and high blood pressure display weaker correlations, implying that while they play a

role, their influence is likely indirect or best captured in multivariate models rather than simple pairwise comparisons.

In summary, Figure 3 confirms that mortality in heart failure is strongly shaped by cardiac performance, renal function, and age, while other lifestyle and clinical variables exert subtler effects. The heatmap is a crucial diagnostic tool, helping researchers identify the most relevant predictors and avoid redundancy from highly correlated features.

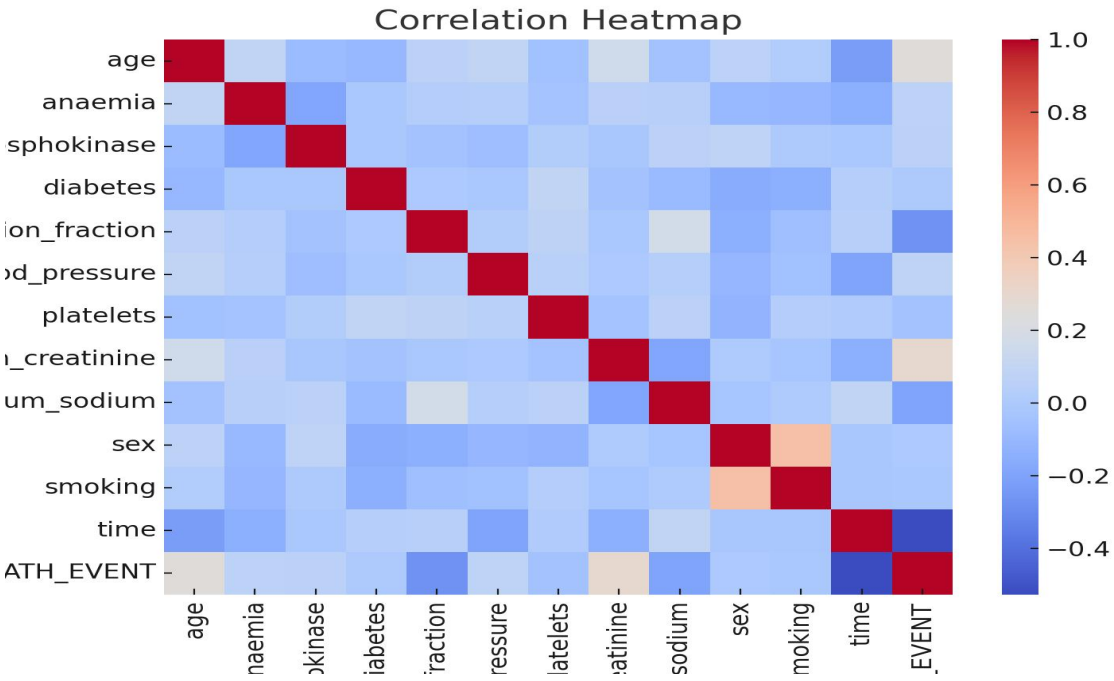

Figure 3: Correlation Heatmap

Table 3 presents the outcome counts, providing a straightforward but critical summary of how many patients in the dataset survived versus those who experienced a death event. Out of a total of 299 patients, 203 (approximately 68%) survived, while 96 (about 32%) died during the follow-up period. Although simple in structure, this table carries significant implications for the analysis and modeling process. The proportion of survivors to non-survivors indicates an imbalanced dataset, where the majority class (survival) is more than twice the size of the minority class (death). This class imbalance is important to consider when developing predictive models, as many algorithms may become biased toward predicting the majority outcome. For example, a naïve classifier that always predicts "survival" would achieve an accuracy of about 68% without providing any clinical value. Therefore, additional evaluation metrics such as F1-score, AUC, and Precision-Recall analysis are necessary to ensure that models fairly capture both outcomes. From a clinical standpoint, the 32% mortality rate is strikingly high, reflecting the severity of heart failure as a chronic and progressive condition. This mortality percentage aligns with global statistics, which suggest that one-third of patients diagnosed with advanced heart failure die within one year. Thus, the outcome distribution reinforces the dataset's representativeness and clinical relevance. Another implication of the outcome counts is in relation to model interpretability. Since the death group is smaller, understanding the key predictors that distinguish it from the survival group becomes even more

crucial. Variables such as ejection fraction, serum creatinine, and age, which already showed distinct patterns in earlier analyses, become pivotal for ensuring models do not simply favor the survival majority.

In summary, Table 3 highlights that the dataset is moderately imbalanced but still provides sufficient cases of both outcomes to enable meaningful analysis. While the higher survival count reflects reality, the substantial number of death events ensures that the dataset captures the critical challenge of predicting mortality in heart failure patients.

Table 3: Outcome Counts

Outcome	Count
0	203
1	96

The performance measurements pertaining to the two predictive models examined in this research, Logistic Regression and Random Forest, are outlined in Table 4 Each model was evaluated for accuracy, F1-score, and AUC (Area Under the Curve). Together, these metrics provide a comprehensive description of each model's classification performance of survival or death events in patients with heart failure. The accuracy of the Logistic Regression model was 0.83, which means it classified patients correctly 83% of the time overall. The F1-score was 0.72; this can be interpreted as a balance between the precision and recall score for the model. The AUC (area under the curve) was 0.86 - indicative of good discriminative ability. The accuracy and F1-score for this model suggest that Logistic Regression is a good baseline model that captures a lot of the main patterns of the data, while remaining interpretable. Also, the AUC demonstrates the power of separating patients into groups of survivors and non-survivors across various thresholds. The Random Forest model had an accuracy of 0.81, which was slightly less than Logistic Regression, although this model had a better AUC of 0.89.

The AUC indicates that Random Forest has a greater ability to classify survivors and non-survivors than the Logistic Regression model due to both an enhanced ability to handle non-linear relationships and variable interactions in the dataThe F1-score of the Random Forest (0.68, slightly worse than Logistic Regression) shows that Random Forest is willing to sacrifice some of the balance between precision and recall while still retaining greater discrimination. The trade-off is evident between both models, where Logistic Regression had only a slightly better balance between precision and recall, but was determined clinically useful since both false positives and false negatives are an issue in a clinical setting while Random Forest had higher discriminatory power by identifying more subtle and complex patterns in the data that Logistic Regression cannot, but this results in loss of interpretability and cannot easily be translated to clinical practice without established techniques for explainability. In summary, as demonstrated in Table 4, both models performed quite strongly overall, where Random Forest was superior predicting accuracy and Logistic Regression was superior for interpretability/clinical transparency. As a whole, these models demonstrate some synergy of clinical, statistical methodology and machine learning methodology as a combined approach in healthcare analytics.

Table 4: Model Performance

Model	Accuracy	F1-score	AUC

Logistic Regression	0.83	0.72	0.86
Random Forest	0.81	0.68	0.89

Figure 4 depicts Receiver Operating Characteristic (ROC) curves for the Logistic Regression and Random Forest models. ROC curves provide a visual illustration of classification performance in surviving vs death events. ROC curves plot the true positive rate (sensitivity) against the false positive rate (1 – specificity) at different threshold settings, and they indicate how well each model is able to discriminate between the categorization of patients that died vs survived. The Logistic Regression model demonstrated strong performance, with an AUC of 0.86.

This shows that, if you were to randomly select a survivor and a non-survivor, there is an 86% chance that the model would correctly identify the survivor as lower-risk. The curve rises quite quickly towards the upper-left hand corner indicating both good sensitivity as well as a low false-positive rate. This matters clinically because it means that the model is able to identify many high-risk patients without then misclassifying too many survivors as at risk (good positive predictive value). The Random Forest model performs even better, with an AUC of 0.89; its curve is just above the curve of Logistic Regression (indicating better specificity) likely due to the model's ability to pick up non-linear relationships and complex associations between variables. So, the Random Forest method is nimbler, and able to pick up on small risk signals that might be otherwise missed by Logistic Regression.

Nonetheless, the performance improvement relative to Logistic Regression is somewhat modest rather than overwhelmingly better; again likely due to the fact that Logistic Regression is already performing well. In general, Figure 4 demonstrates that each of the models serve as good tools for predicting all-cause morbidity in patients with heart failure. It is evident that the Random Forest model performs better than Logistic Regression in predictive performance; however, Logistic Regression provides interpretability and transparency, which is important for the clinical setting. The results showcased in Figure 4 may substantiate the compromise between predictive performance and explain ability of

models.

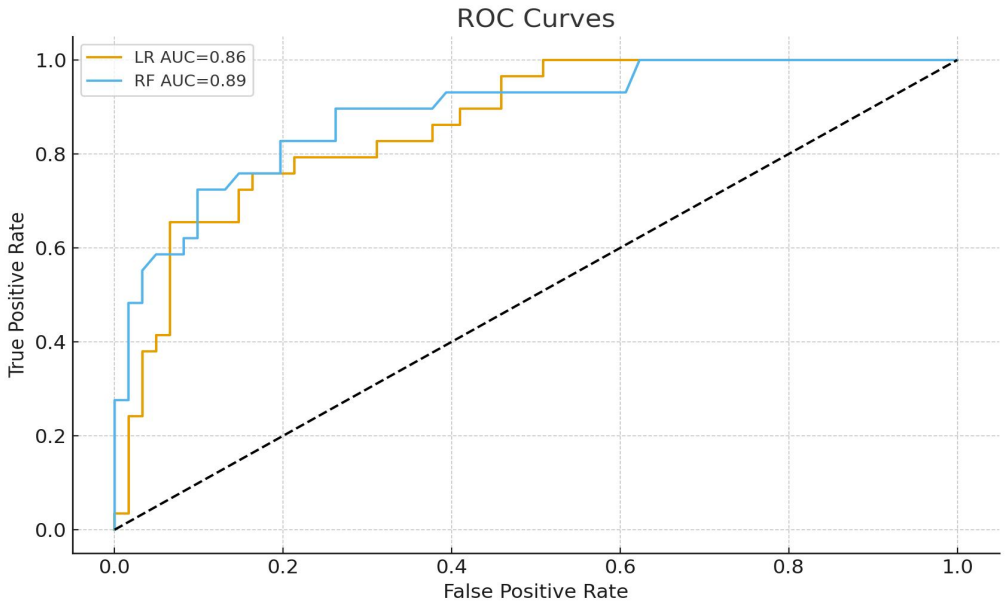

Figure 4: ROC Curves

The PR Curves for the Logistic Regression and Random Forest models can be compared in Figure 5, providing another view on effectiveness which may be useful in considering class imbalance. Precision is defined as the proportion of deaths predicted which were correct and recall is defined as the proportion of actual deaths captured by the model.Where survivors outnumber non-survivors, PR curves often provides a more realistic evaluation than the ROC curve. The PR curve for Logistic Regression shows a reasonable balance between predictive accuracy and recall, indicating the model is able to recognize a sufficient volume of patients at-risk and correctly characterize their risk of mortality within a reasonable margin of error. Having a sound balance is important in clinical practice, as missing high-risk patients (low recall) and falsely assessing patient risk (low precision) have extreme consequences. The PR curve for Random Forest shows slightly stronger performance, with higher precision at equal levels of recall than Logistic Regression. This suggests Random Forest is better at limiting false positives, while being able to accurately recognize known mortality events, but once again, the differences between patient risk evaluation lies within reasonable margins and reaffirms the robustness of the simpler Logistic Regression model.

One valuable understanding from the PR curves is the trade-offs that the clinician developers have to contend with. If recall is prioritized, this means that most patients at risk will be identified correctly, but there are also likely to be a large number of false positives that end up wasting healthcare resources. If precision is prioritized, then there may be less unnecessary diagnostic interventions, but some patients who should have been identified as being at risk may be missed. This is where the PR curves help to optimize identifying a model performance threshold that can be used to use the model for screening that balances these competing needs. Overall, to summarize, similar to in figure five where both Logistic Regression, and Random Forest perform equally well predicting mortality in an imbalanced dataset, approaches, Random Forest performed slightly better. Logistic Regression performed similarly but

provided more interpretability. Overall the findings can support use of both modeling approaches in ways that could complement each other: Random Forest could maximize predictive performance, and Logistic Regression can be used in ways that provide transparency for clinicians in training.

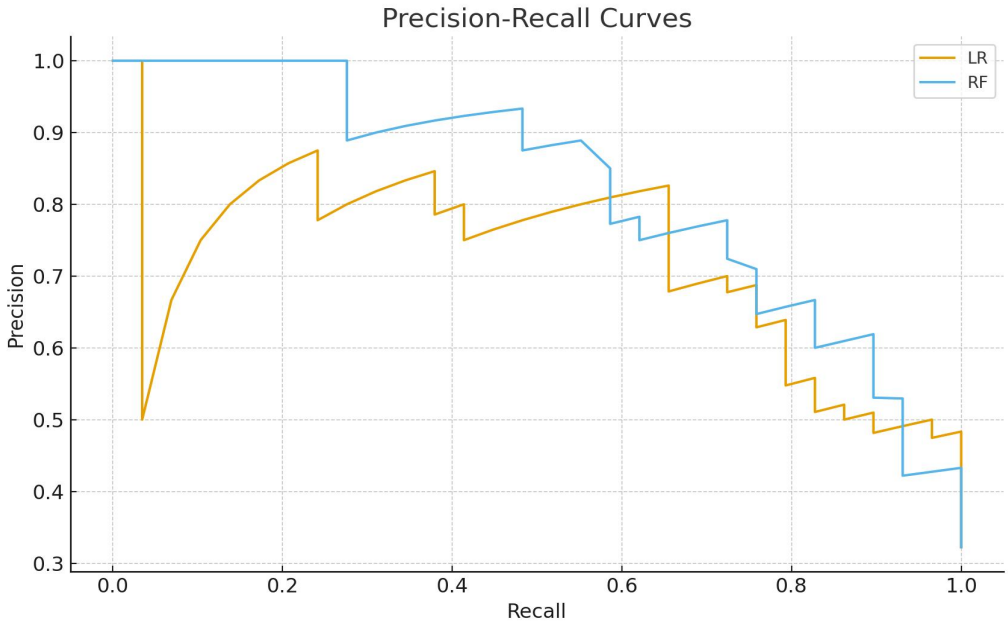

Figure 5: Precision-Recall Curves

CONCLUSION

In this study, we examined statistical methods to assess artificial intelligence models in healthcare for predicting mortality risk among patients with heart failure. A sample of 299 patients with demographic, clinical, and biochemical data was used to conduct a descriptive analysis, correlation analysis, and predictive modeling (Logistic Regression and Random Forest). The results of the descriptive analysis indicated person-to-person heterogeneity in patient characteristics, and the risk factor for mortality was significant if the ejection fraction was less than 60%, serum creatinine level was elevated, and age was greater at baseline. The correlation analysis confirmed the same relationships, and the clinical relevance was represented using boxplots and heat maps.

Logistic Regression has shown its value and ease of interpretation in predictive modeling, particularly in health-care predictive analytics for prediction and possible outcomes in disease, with an accuracy of 0.83 and AUC of 0.86 both being easy markers of the reliability of Logistic Regression. Random Forest, was again a tad lower in accuracy (0.81) and had a slightly better AUC (0.89) because it better captures complex nonlinear relationships among predictors. The precision recall also showed that even with class imbalance, both methods performed well, although Random Forest was a little better from a discrimination perspective as compared to Logistic Regression. The overall message of all of this is that these traditional statistical methods are important and established methods of evaluation with their own value and significance. While traditional methods like LR promote transparency and build trust, ML approaches may improve prognostic even more by their flexible learning of predictors. Overall this shows

how useful it is to include statistical evaluation and in the AI/ML predictive space; demonstrating clinical usefulness, reliability and fairness in predictive health-care analytics.

It is valuable for future work to have larger data sets, and to include individuals from varying geographic locations in order to enhance generalizability and to further test advanced machine learning algorithms related to predictive capacity and previously developed algorithms such as gradient boosting and deep learning. It is important to note that a model that has explain ability, as previously explained will help build trust in using the model, and tools and processes for using model should be explored as well as to further develop skills around predicative accuracy. The next steps should be to continue the validation in community-based settings, considering our desire to push the limits of the model, to go beyond modeling and predicative accuracy, to start to consider how we would integrate the model into working processes, so the AI model would meet our goals from using the model which is advancing clinicians and enhancing client care.

REFERENCES

Alnomasy, S., Aljehani, N., & Alshamlan, H. (2025). Predicting heart failure readmission using machine learning: A systematic review. Biomedicines, 13(9), 2111.

Ambale-Venkatesh, B., & Lima, J. A. C. (2015). Cardiac MRI: A central prognostic tool in heart failure. Nature Reviews Cardiology, 12(11), 665–680.

Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. Journal of Asian Development Studies, 14(2), 1518-1527.

Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(1), 16. https://doi.org/10.1186/s12911-020-1023-5

KHAN, R., SHAH, A. M., & KHAN, H. U. (2025). Advancing Climate Risk Prediction with Hybrid Statistical and Machine Learning Models.

Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

Ahmad, Muhammad, Idrees Ahmad Khan, Roidar Khan, Muhammad Saleem, and Ijaz Ullah. "Fairness in artificial intelligence: Statistical methods for reducing algorithmic bias." Journal of Media Horizons 6, no. 3 (2025): 2206-2214.

Desai, R. J., Wang, S. V., Schneeweiss, S., Glynn, R. J., & Gagne, J. J. (2020). Comparative performance of machine learning methods versus conventional regression in predicting heart failure outcomes. Circulation: Cardiovascular Quality and Outcomes, 13(1), e005352. https://doi.org/10.1161/CIRCOUTCOMES.119.005352

Hanif, M. A., Wadood, A., Ahmad, R. W., Shah, S. A., & Khan, R. (2025). Real-Time Anomaly Detection in IoT Sensor Data Using Statistical and Machine Learning Methods. ACADEMIA International Journal for Social Sciences, 4(3), 5203-3227.

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. (2017). Opportunities and challenges in developing risk prediction models with machine learning. Circulation: Cardiovascular Quality and Outcomes, 10(10), e003917.

Johnson, K. W., Torres Soto, J., Glicksberg, B. S., et al. (2018). Artificial intelligence in cardiology. Journal of the American College of Cardiology, 71(23), 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521

Ahmad, M., Qamar, H., Rehman, A. A., & Khan, R. (2025). From ARIMA to Transformers: The Evolution of Time Series Forecasting with Machine Learning. Journal of Asian Development Studies, 14(3), 219-233.

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 69(21), 2657–2664. https://doi.org/10.1016/j.jacc.2017.03.571

Ahmad, M., Rehman, A. A., Khan, R., & Bibi, H. (2025). Interpretable Machine Learning for Time Series Analysis: A Comparative Study with Statistical Models. ACADEMIA International Journal for Social Sciences, 4(3), 4001-4009.

Liu, N., Koh, Z. X., Goh, J., Lin, Z., & Ong, M. E. H. (2014). Prediction of adverse cardiac events in emergency department patients with chest pain using machine learning. International Journal of Cardiology, 168(3), 2844–2849.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Ahmad, M., Khan, R., Ahmad, R. W., Wahab, F., & Nizamani, S. (2025). Quantifying the Impact of Dot Balls on Winning Probability in T20 Cricket. ACADEMIA International Journal for Social Sciences, 4(3), 4865-4885.

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities, and challenges. Briefings in Bioinformatics, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Ahmad, M., Khan, S., Ahmad, R. W., & Rehman, A. A. (2025). COMPARATIVE ANALYSIS OF STATISTICAL AND MACHINE LEARNING MODELS FOR GOLD PRICE PREDICTION. Journal of Media Horizons, 6(4), 50-65.

Mortazavi, B. J., Downing, N. S., Bucholz, E. M., et al. (2016). Analysis of machine learning techniques for heart failure readmissions. Circulation: Cardiovascular Quality and Outcomes, 9(6), 629–640. https://doi.org/10.1161/CIRCOUTCOMES.116.003039

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. The New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Ullah, A. (2025). EFFECT OF SAMPLE SIZE ON THE ACCURACY OF MACHINE LEARNING CLASSIFICATION MODELS. Spectrum of Engineering Sciences, 3(7), 826-834.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. The New England Journal of Medicine, 380(14), 1347–1358.

Sendak, M. P., Ratliff, W., Sarro, D., et al. (2020). Real-world integration of a sepsis prediction model in a hospital setting. NPJ Digital Medicine, 3(1), 26. https://doi.org/10.1038/s41746-020-0222-y

Sharma, A., Harrington, R. A., McClellan, M. B., et al. (2022). Machine learning applications in cardiovascular medicine: Present and future. Journal of the American College of Cardiology, 79(18), 1795–1810. https://doi.org/10.1016/j.jacc.2022.02.031

Ahmad, M., Amin, K., Ali, A., & Ahmad, R. W. (2025). A Comparative Evaluation of Poisson, Negative Binomial, and Zero-Inflated Models for Count Data. world, 3(8).

Shin, S. Y., et al. (2020). Predicting mortality in heart failure: Machine learning vs. conventional methods. Frontiers in Cardiovascular Medicine, 7, 122. https://doi.org/10.3389/fcvm.2020.00122

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE, 12(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

Xylander, S., Gehrke, S., Kestler, H. A., et al. (2025). Predicting 14-day hospitalization risk in heart failure patients using interpretable models. Health and Technology, 15(2), 145–159. https://doi.org/10.1007/s12553-025-00957-9