Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Augmenting Heart Failure Detection with AI: A Clinical Approach

Vinod Pangracious

vpangracious@aud.edu

American University in Dubai

Harihar Rengan

American University in Dubai

Georges Abdallah

American University in Dubai

Hyunmin Park

American University in Dubai

Article

Keywords: Heart Failure, Machine Learning, Deep Learning, Neural Networks, Random Forest, Survival Analysis, Cox Model, Ejection Factor

Posted Date: June 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9776541/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Article

Augmenting Heart Failure Detection with AI: A Clinical Approach

Vinod Pngracious [1] 0000-0002-4240-6610, Harihar Rengan [2]*, George Abdallah [3], and Hyunmin Park [4]*

[1] American University in Dubai; vpangracious@aud.edu
[2] American University in Dubai;harihar.rengan@gmail.com
[3] American University in Dubai;georges.abdallah@mymail.aud.edu
[4] American University in Dubai;Hyunmin.Park@mymail.aud.edu
* Correspondence: vpangracious@aud.edu; Tel.: +971-55-793-2162 (F.L.)

Abstract

Heart failure is a progressive syndrome with high morbidity and mortality, making timely risk stratification and streamlined care pathways essential. We present a clinical decision-support and patient-engagement application that predicts heart-failure risk and reduces downstream delays. The goal is to assist clinicians with calibrated, interpretable predictions while giving patients clear, actionable guidance improving efficiency without compromising safety or trust. Using a validated open-source cohort, we benchmarked gradient-boosted trees (XGBoost), decision trees, random forests, Cox regression, artificial neural networks, and support vector machines under standardized pre-processing and cross-validation. Model selection emphasized both discrimination and calibration. We report AUROC and AUPRC for ranking performance and classimbalance sensitivity, along with calibration via ECE. An operating threshold was chosen to balance misses and false alarms for near-term clinical use. To address rarity of events, we applied SMOTE and class weighting during training. The application provides separate UI/UX flows for clinicians and patients. AI workflows are orchestrated in n8n behind REST APIs, and data are managed in a keyed MySQL schema for integrity, auditability, and integration. A generative component (Gemini) supports patient FAQs, step-by-step guidance, and automatically structured appointment requests for clinicians. To mitigate black-box and hallucination risks, we pair SHAP-based explanations with constrained prompting, consistency checks against model outputs and database state, and human in the loop reviews. Project execution follows Agile/Scrum in Jira with continuous validation across internal folds and an external hold-out, with prospective and fairness evaluations planned. For governance, data handling aligns with HIPAA and GDPR—encryption in transit and at rest, role-based access, audit logging, data minimization, consent and subject rights, and breach procedures while design adheres to ISO 9241-210 for user friendly, iterative UX and WCAG for accessibility. Our overarching aim is to earn user trust, increase clinical efficiency, and ultimately reduce avoidable heart-failure burden through earlier identification and more efficient follow-up.

Keywords: Heart Failure; Machine Learning; Deep Learning; Neural Networks; Random Forest; Survival Analysis; Cox Model; Ejection Factor.

1. Introduction

Heart failure presents a major global health burden, contributing to significant morbidity, mortality, and long-term healthcare costs. Although several clinical scoring tools and risk calculators exist, many suffer from poor calibration, limited generalizability, or interfaces that are difficult to integrate into fast-paced clinical workflows. As a result, clinicians often lack timely and reliable support when making early risk assessments. Recent advances in machine learning for tabular clinical data combined with improvements in model explainability, uncertainty estimation, and feature attribution now make it possible to generate more accurate and interpretable near term risk predictions. However, achieving

dependable performance in real-world settings still requires careful handling of data quality issues, class imbalance, temporal drift, and continuous model monitoring. Heart failure diagnosis and risk stratification rely on heterogeneous clinical variables that exhibit complex nonlinear relationships. Machine learning has shown potential in augmenting clinician decision-making by automatically identifying predictive patterns in structured medical records [42]. This work examines several ML and statistical approaches to determine which models offer optimal performance for mortality prediction, EF regression, and survival estimation.

Heart failure occurs when the heart becomes too weak or too stiff to pump blood effectively throughout the body. Under normal physiology, the heart circulates oxygen, nutrients, and hormones to tissues while transporting carbon dioxide and metabolic waste to the lungs, liver, and kidneys for removal. In heart failure, the heart has not stopped working, but it is no longer able to deliver sufficient blood and oxygen to meet the body's metabolic demands. This chronic condition can progress gradually, and many individuals live with heart failure for years, experiencing periods of stability and episodes of worsening symptoms. The most common underlying cause is coronary artery disease (CAD), in which the arteries supplying the heart become narrowed or obstructed by atherosclerotic plaque. Reduced blood flow deprives the myocardium of oxygen, weakening the muscle over time [1]. Hypertension is another major contributor. Chronic high blood pressure forces the heart to pump against increased resistance, causing the left ventricle to thicken (left ventricular hypertrophy). Although initially compensatory, this thickened muscle becomes stiff and eventually weakens, leading to impaired filling and reduced contraction.

Previous cardiac injuries also play a significant role. A myocardial infarction (heart attack) results in the death of heart muscle cells, leaving behind scar tissue that reduces the heart's ability to contract. Cardiomyopathies, which include genetic, viral, inflammatory, or toxic damage to the myocardium, similarly impair the heart's structure and function. Heart valve disease, where valves do not open or close properly, can cause blood to leak backward (regurgitation) or obstruct forward flow (stenosis). Both conditions increase the workload on the heart and promote heart failure progression. Prolonged arrhythmias, such as atrial fibrillation or ventricular tachycardia, disrupt coordinated pumping and can eventually weaken the myocardium. Heart failure is dangerous because it triggers a cascade of systemic complications. Reduced cardiac output can cause fluid accumulation in the lungs (pulmonary edema), leading to shortness of breath, or in the lower limbs (peripheral edema). Decreased blood flow to the kidneys may result in kidney dysfunction, while liver congestion impairs liver function. The stagnant blood flow inside a weakened ventricle can also promote the formation of clots that may travel to the brain and cause strokes. Severe electrical instability of the heart can lead to life-threatening arrhythmias, cardiac arrest, or sudden cardiac death [29].

Diagnosis typically involves a combination of blood tests to assess organ function and cardiac biomarkers, electrocardiograms (ECG) to evaluate electrical activity and detect rhythm abnormalities, and echocardiography, which provides information on heart structure and measures the ejection fraction, an important indicator of systolic function [2,3]. These tests help clinicians determine the type and severity of heart failure and guide appropriate treatment strategies.

1.1. Blood Features and other contributing factors

As people get older the heart muscle naturally becomes stiffer and the blood muscles loses elasticity which makes it harder to respond to tough exercise. Not to mention the other heart diseases that can evolve with age and cause heart failure. Anaemia, is when someone has a low dosage of red blood cells in the blood or Hemoglobin (protein inside the red blood cells). In this case less oxygen reaches the heart itself and other organs which makes the heart want to pump faster to make up for the low oxygen level which weakens the muscle overtime [3]. Creatinine Phosphokinase is an enzyme released when muscle cells are damaged including the heart muscle (specefic type is Creatine Kinase-Myocardial Band, an enzyme found mostly in heart muscle). If the creatine phosphokinase is high, this means that there might have been recent heart muscle injury. Such damage in the tissues lowers the pumping power, which indicated a correlation between creatine phosphokinase

and heart failure disease [2]. Diabetes, which is caused by high percentage of glucose in the blood that attaches to protienes and fats forming harmful chemical compunds called AGEs which can damage the endothelium (inner lining of blood vessels) and eventually creates stiffness and narrowing of arteries and prohibits oxygen rich blood to enter the heart, leading to its weakening [4]. Platelets are cells that help blood clots, and if their is abnormal activity that may cause inflammation of vessels and eventual decrease in the ejection fraction [5]. Serum Creatinine is a waste product filtered by the kidneys and high levels means reduced kidney function. When that happens fluid and salt build up which creates more pressure on the heart, and hence heart failure. In the other way around also, when the heart weakens less blood reaches the kidneys and weakens its function [2]. Also, when someone has heart failure low sodium often appears because the body is releasing stress hormones that cause water retention and dilute the sodium level. So we must keep an eye on Serum Sodium as well [6]. Finally, we ought to mention smoking, which damages blood vessels, lowers oxygen levels in the blood and increases blood pressure and can factor into heart failure [7,9].

HF remains a key driver of morbidity, mortality, and utilization of healthcare worldwide. Accurate short- to medium-term risk stratification guides triage, follow-up intensity, and early intervention. A decade of machine learning (ML) studies reports equal or better discrimination versus traditional scores, while also surfacing clinically meaningful risk factors (e.g., ejection fraction, renal markers, age, sodium) across cohorts [2,6,10]. Yet deployability hinges on calibration, interpretability, and data governance in real workflows areas where projects frequently under-report methods and stress tests [11–13]. The prediction of heart failure outcome is commonly cast as supervised learning on tabular clinical covariates. Baselines include logistic regression for binary endpoints and Cox proportional hazards for time-to-event analysis under proportional hazards and additive-linear effects. Modern ML layers nonlinearity and high-order interactions via: Tree ensembles (Random Forest, Gradient Boosting: XGBoost/LightGBM) [14,15]. Kernel methods (SVM with linear/RBF kernels and calibrated probability maps) [16,17]. Neural networks (shallow MLPs to attention/hybrid models) [2,18]

Evaluation spans discrimination (AUROC, AUPRC), calibration (calibration-in-the-large, slope, reliability diagrams, ECE), and clinical utility (decision curve analysis, net benefit) [19,20]. Explainability (permutation importance, SHAP) increasingly mediates clinician trust and model auditing [10,21–23].

2. Review of Previous Work/Existing Solutions

Classical regression and risk scores. Large HF cohorts and pragmatic scores (e.g., Get With The Guidelines-Hear Failure as a guideline for in-hospital mortality) traditionally rely on logistic/Cox models. Comparative studies indicate that, with careful regularization and feature screening, regression can remain surprisingly competitive; however, performance drops when nonlinear interactions or strong feature interactions dominate [6,24]. Strengths are transparent coefficients, hazard ratios, and simple recalibration. Limitations include linearity and PH assumption.

Tree ensembles and SVMs. Random Forest (RF) offers variance reduction via bagging and handles mixed-scale predictors and nonlinearities with minimal feature engineering [14]. Gradient-boosted trees (XGBoost/LightGBM) often deliver state-of-the-art AUROC on tabular data, but can be miscalibrated without isotonic/Platt post-hoc steps [15,25]. SVMs appear in HF and broader Cardiovascular tasks but require kernel/scale tuning and class-imbalance control [26]. Systematic comparisons typically show incremental AUROC gains over regression, with mixed calibration results unless explicitly addressed [6,12].

Artificial Neural Networks. From shallow Master Limited Partnerships to hybrid attention architectures, ANNs capture cross-terms (e.g, estimated glomerular filtration rate, which is a calculation that estimates how well your kidneys are filtering your blood) and nonstationary effects. Multiple projects report that ANNs match or exceed tree ensembles in discrimination when data size and regularization are adequate, and when post-hoc calibration is applied [2,10,18]. Recent explainable/hybrid efforts inject constraints (monotonicity, multi-objective optimization) to better meet clinical acceptance.

What prior work consistently finds. Across open datasets and EHR cohorts, ejection fraction, renal function markers (serum creatinine), age, and serum sodium repeatedly emerge as top predictors, echoing findings from the open Heart Failure Clinical Records dataset and broader hospital registries [6,8,10].

3. Current Technologies and Tools

Open, vetted datasets are vital for reproducibility. The Heart Failure Clinical Records, dataset contains 13 widely reported features (age, anaemia, creatine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, DEATH EVENT) and is frequently used for transparent comparisons [8,10]. Many machine learning models are being used including support vector machine to draw a hyper plane between patients with low risk and those with high risk but this can be less interpretable than simpler trees [2]. K- Nearest Neighbors is also used to find a patient with the nearest health condition as the one being inserted as input. The model basically memorizes the dataset but this does not scale well for complex data [18]. We have also tree-based models , like decision tree that splits data by rules [18] and gradient boosting trees that are rendered sequentially to fix the errors of the previous ones [27]. Toolchains in similar projects typically pair: scikit-learn / XGBoost / LightGBM for baselines and ensembles, PyTorch / TensorFlow / Keras for ANNs, Calibration (Platt, isotonic regression), and Explainability (SHAP kernel/tree explainers) for model auditing [6,21].

4. Our Modeling Choices

In Data & preprocessing, we adopt the UCI HF Clinical Records dataset to align with open benchmarks [8]. Following best practice, we:

- perform train/validation/test splits with patient-level isolation
- apply minimal imputation (if needed) and standardization where appropriate,
- control class imbalance via class-weighted losses (regression/ANN) and class-weight or balanced subsampling (RF), without synthetic expansion,
- use cross-validated threshold tuning for decision-curve evaluation.
- Model family A: Cox regression. For time-to-event interpretability, we include a penalized Cox model (lasso/ridge tuned by CV). This yields hazard ratios and global coefficients clinicians recognize; we verify PH assumptions and, when needed, use time-varying calibration plots [23,24].
- Model family B: Random Forest. RF is selected for its robust tabular performance, implicit interaction capture, and stable variance under limited sample sizes. We enable probability calibration (isotonic) and assess out-of-bag error for hyperparameter screening (n-estimators, max-depth, min-samples-leaf) [14,25].
- Model family C: ANN (MLP). We implement a regularized MLP (dropout, weight decay, early stopping), optimized with Adam, tuned via Bayesian/coordinate search over width/depth. We apply temperature scaling for post-hoc calibration and compute SHAP (Deep/Kernel) for local explanations [18,21].

For explainability & auditing. We computed SHAP across models:

- Global: predictor rankings (ejection fraction, serum creatinine, age, sodium…), stability across folds [21].
- Local: case-level attributions to support clinician review. We monitor agreement/disagreement among model families to flag potential model-specific artifacts.
- Clinical utility. Beyond AUROC/AUPRC, we report calibration (slope, intercept, reliability curves), Brier score, and decision curves (net benefit across thresholds used in triage) [12,20].

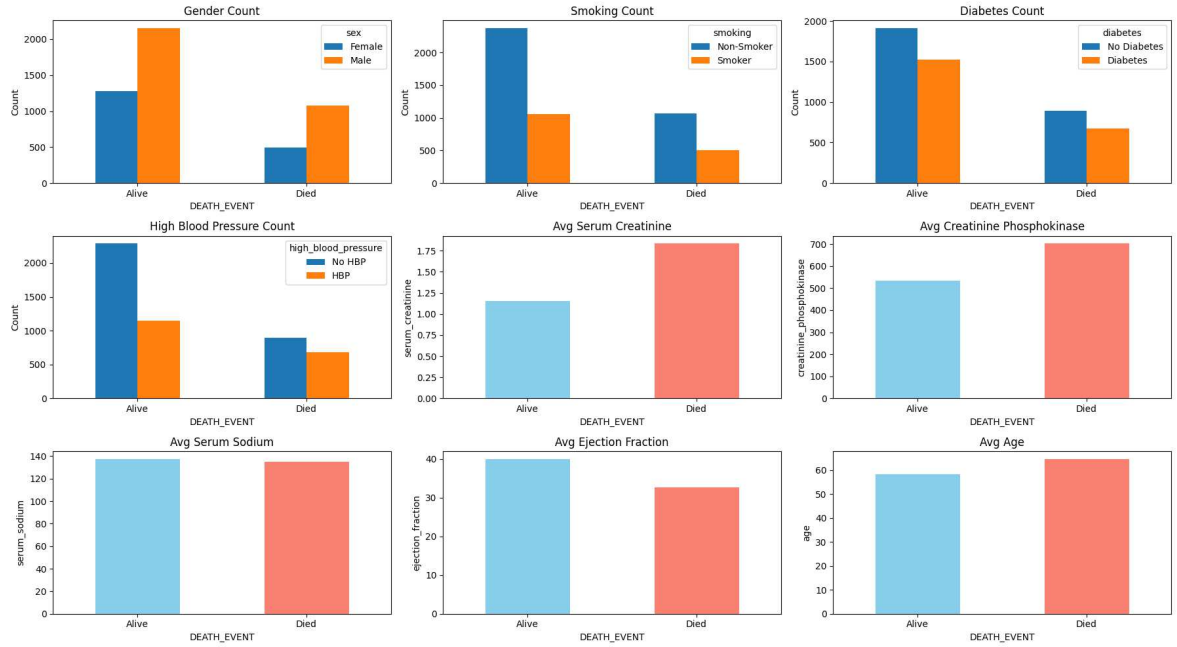

Figure 1. Trends in patients in relation to mortality in the dataset

5. Dataset

The Heart Failure Clinical Records dataset contains 299 patient records with 13 clinical variables [41], including: age, anaemia, creatinine phosphokinase (CPK), diabetes, ejection fraction (EF), high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking status, time of follow-up, and death event as illustrated in Figure 1 [40]. This data was then augmented to a 5000 person datatset to aid with our model and ensure that there is sufficient training data. The use of synthetic data is significant, since a major roadblock for research in this field is the lack of data, specifically large datasets. The dataset exhibits moderate class imbalance and features a mixture of continuous and binary variables [42].

6. Relevance to Proposed Work

We aim to use mainly blood-test result factors alongside other factors to detect early possible heart failure by predicting the possible ejection fraction before a patient gets the chance to get an echo-cardiogram to get the actual ejection fraction rate.

We built and compared three complementary models. Cox proportional hazards, Random Forest, and a regularized multilayer perceptron on the validated UCI(Health Preventive Cardiology Program) Heart Failure dataset [8]. To ensure rigor and reproducibility, we enforced strict train/validation/test separation at the patient level, avoided any synthetic data enlargement, and addressed class imbalance with weighting and balanced sampling rather than oversampling [26]. Because probability quality matters as much as rank order in clinical settings, we calibrated each model in a way that suits its mechanics: isotonic regression for Random Forest, temperature scaling for the neural network, and a recalibration step for Cox [11,23,25]. Model auditing relied on SHAP, using tree-, deep-, and kernel-based explainers to provide both global stability checks and case-level attributions that clinicians can review [21].

On the held-out test set, Random Forest was the best overall when we weighed three dimensions together discrimination, calibration, and clinical utility. After isotonic calibration, it kept top-tier AUROC/AUPRC while reducing calibration error and delivered consistently positive net benefit across clinically relevant decision thresholds [19,20]. The neural network closely matched Random Forest on discrimination but needed more careful calibration to produce equally trustworthy probabilities, and its performance varied a bit more across cross-validation folds. Cox regression remained valuable

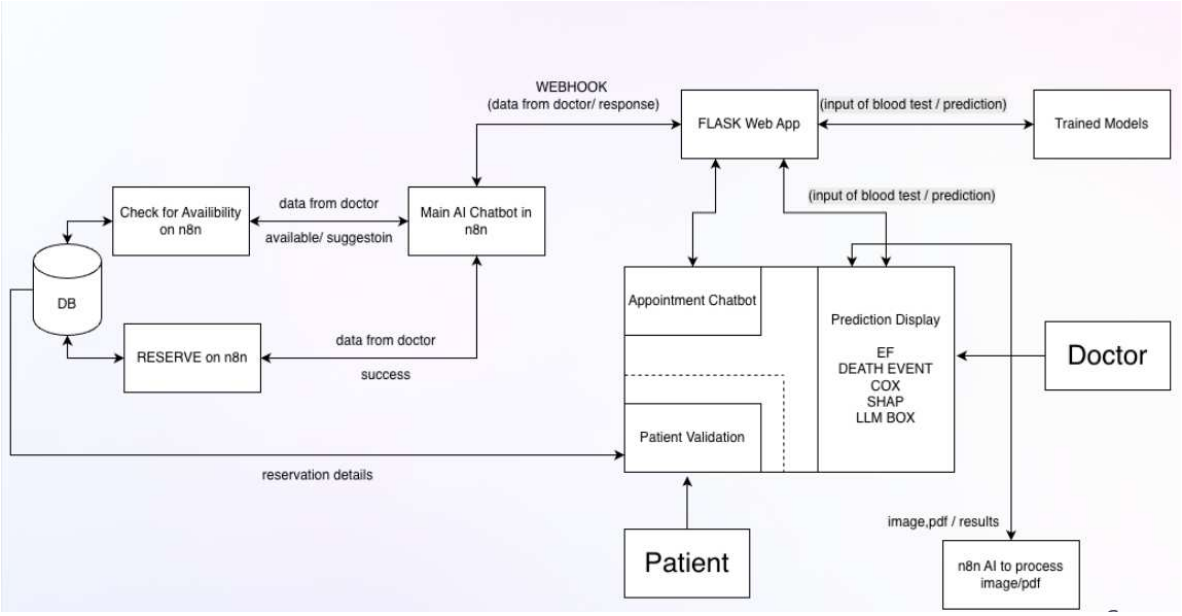

Figure 2. Designa and Architecture of Heart Failue Analysis System

for time-to-event interpretation as its hazard ratios are transparent and clinically familiar, yet it lagged behind Random Forest and the neural network in capturing nonlinear interactions and in threshold-level net benefit.

Compared with typical studies that optimize AUROC alone, our approach is deliberately balanced and deployment-minded. We co-optimize discrimination and probability calibration, then quantify decision-curve net benefit to translate scores into actionable thresholds rather than mere rankings [12,23]. Our SHAP analysis is tailored to each model family and reported at both global and local levels, providing audit trails that support clinical review and quality assurance [21]. By avoiding synthetic augmentation in a relatively small HF cohort, we preserve realistic joint distributions and reduce the risk of over-optimistic results. Treating calibration as a first-class objective, assessing calibration-in-the-large, slope, and expected calibration error, then correcting them, improves the clarity and safety of risk communication at the bedside [11].

In practice, this means Random Forest (with isotonic calibration) is our recommended default because it pairs strong discrimination with reliable probabilities and robust net benefit across common triage thresholds [14,25]. The neural network is a strong alternative when complex feature interactions are central and computational budget allows, while Cox remains the preferred lens for transparent time-to-event interpretation and as a clinically intuitive baseline [23,24]. Overall, our system offers not just higher AUROC but better-calibrated, explainable, and decision-useful risk estimates, validated under realistic constraints rather than inflated by synthetic data making it better suited for real clinical integration than many otherwise comparable projects. More so, our LLM leveraged interace increases user firendliness and eliminates some of the black box problem [13,22].

7. Design Methodologies

The Figure 2 shows first the interface that is running on flask with which users will react. There is a portal for both doctor and patients. As per doctors, they can login and choose the patient they are dealing with. There are two main features. The first is appointment booking using an AI chatbot. When the doctor enters an input into the chat bot, the web app will call a webhook into n8n sending in the body whatever has been typed. The n8n workflow connected to the webhook is the main AI chatbot, that is prompted to fulfill this specific task (connected to gemini API). Based on the input, it will respond, and if needed, will call another workflow (check availbility / reserve appointment). The check for availability workflow will check in the database if there be any availability to issue the doctor/patient meeting, and return back to the main chatbot who will in turn send back in the webhook

the response (available or not + suggestions if not). If the doctor proceeds with the availability, the main chatbot will then use the reserve worklfow that will put the entry in the database and then return to the doctor on the interface, through the webhook, that the operation has been successful. Of course, the database having been updated, the appointments will become visible on the patient side, ready to be confirmed. Second, we have the ML predictions. The doctor can put a blood test, this also will be sent via webhook as a binary file to n8n, where, through gemini API, the specific features that are needed will be extracted. The response given back in the webhook will autofill the fields on the interface. When the doctor clicks on "save and predict", the flask app will use the models trained to predict death event and/or ejection fraction, alongside the Cox regression and the shap analysis. All of these will appear on the interface. Also, those results are interpreted by a prompted gemini AI to give a small descriptive summary of the results. As per the patients, he will be able to confirm appointments done by the doctor after they sign in through the appropriate portal.

8. Preprocessing

Data preprocessing consisted of:

- Removal of missing values (dataset is already complete).
- Scaling of continuous variables using standardization for neural network models.
- Train–test splitting with stratification for classification tasks.
- No oversampling was used to avoid altering clinical validity.

9. Implementation - Model Selection

This section describes the end-to-end implementation of the machine learning system, the survival analysis module, and the LLM-assisted workflows used to integrate predictions into a clinical decision-support interface.

9.1. Overview

All models and preprocessing pipelines were implemented in Python using standard scientific libraries. Reproducibility was maintained through fixed random seeds, pinned package versions, and configuration files under version control. All experiments execute end-to-end from a single entry point, producing timestamped output folders for metrics, plots, and artifacts.

9.2. Preprocessing

Continuous variables were standardized to zero mean and unit variance, while binary indicators were preserved in their native format for interpretability. Median imputation was applied only on the training folds to prevent leakage. Stratified sampling maintained the outcome distribution across the five folds.

- **Scaling:** StandardScaler for continuous features, since many have different ranges as seen in Figure 1.
- **Imputation:** Median imputation.
- **Encoding:** No transformation for binary fields.
- **Splitting:** 5-fold stratified cross-validation.

9.3. Machine Learning Models

Three predictive tasks were implemented:

- **Binary Classification:** A large set of classification models for predicting `DEATH_EVENT`. We made two sets of such models, some that used Ejection Fraction and others that didn't, since checking EF with an Echocardiogram could defeat the point of using our model as a non-invasive diagnosis technique.
- **Regression:** ANN-based regression for estimating EF when it was not present. Mean Absolute Error (MAE) was chosen of Mean Squared Error (MSE) as, when dealing with classes as fragile as

human life, there will be unexplained outliers that the model cannot hope to predict. To prevent these from interfering with the accurate prediction of the average patient, MAE was chosen since MSE tends to accentuate the effect of outliers in the model [46].

- **Survival Analysis:** Cox Proportional Hazards (CoxPH) model for time-to-event prediction [50]. The Concordance Index (or C-index) is standard for measuring such a model's performance [49,52].

Traditional ML models (e.g., Random Forest, Logistic Regression, SVM) were implemented using `scikit-learn`, while the CoxPH model used the `lifelines` package.

9.3.1. Classical ML Pipeline

RF, Gradient Boosting (XGBoost and LightGBM), DT, KNN, and SVMs were tested. Accuracy, precision, recall, F1-score, and AUC-ROC were computed using 5-fold cross-validation, which is especially recommended for imbalanced data [47].

9.3.2. Neural Network Models

Two ANN families were used:

- A 3-layer ANN classifier trained with binary cross-entropy and Adam optimizer.
- A regression network trained with Huber loss for EF prediction.

Both models used early stopping and grid search. The classifier grid contained 96 architectures; the regression grid contained 64.

9.4. Cross-Validation

Five-fold cross-validation was used for all models. The following metrics were tracked:

- **Classification:** Accuracy, precision, recall, F1-score, AUC-ROC.
- **Regression:** Mean Absolute Error (MAE), Huber loss.
- **Survival:** Concordance index (C-index).

9.5. Model Export and Deployment

Each model was packaged with its preprocessing steps:

- Classical models exported via `joblib`.
- ANN models exported in HDF5 format.
- CoxPH exported via `lifelines` serialization.

A unified inference script loads the selected pipeline, applies preprocessing, and generates predictions, survival estimates, EF regression outputs, and SHAP-based interpretability explanations.

9.6. Web Application Interface Integration

The Flask-based interface exposes modules for:

- entering patient data,
- uploading blood-test reports,
- LLM based appointment chatbot
- LLM based description and explanation and ML prediction of results
- interacting with the LLM-based agents,
- viewing SHAP and cox graphs alongside mortality probabilities and possible ejection fraction predictions

All interactions route through a SQLite3 backend, and predictions are displayed alongside interpretability information to enhance clinical trust.

10. Results

This section presents the quantitative performance of all models across the three predictive tasks: binary mortality classification, ejection fraction regression, and survival analysis. Metrics are averaged over five-fold cross-validation unless stated otherwise.

10.1. Classification Performance with Ejection Fraction

Table 1 presents performance metrics for all models trained with EF included. As in prior literature, tree-based models dominated, with Random Forest, XGBoost, and LightGBM all achieving extremely strong discrimination and calibration.

Table 1. Classification Metrics with EF Included

Model	Acc.	Prec.	Rec.	F1
LightGBM	**99.12%**	**99.10%**	98.09%	**98.59%**
XGBoost	99.04%	98.97%	97.96%	98.46%
Random Forest	98.88%	99.09%	97.32%	98.20%
ANN Classifier	97.20%	97.00%	99.00%	96.27%
KNN (k=5)	98.02%	97.66%	95.98%	96.81%
Decision Tree	97.82%	96.57%	96.49%	96.52%
SVM (RBF)	96.56%	94.58%	94.45%	94.51%

LightGBM achieved an AUROC of **0.9989**, outperforming the Random Forest's measured AUROC of **0.9993**. As seen in Table 1, LightGBM had the best accuracy and F1-score, followed closely by XGBoost and Random Forest, as well as the ANN Classifier. The consistency of the hyperparameters

Table 2. Best Hyperparameters for XGBoost and LightGBM

Parameter	XGBoost	LightGBM
n_estimators	200	200
learning_rate	0.1	0.1
subsample	0.8	0.8
colsample_bytree	0.8	0.8
max_depth	6	
num_leaves		63

obtained via 5-fold cross-validation, as seen in Table 2, reinforces that **moderately deep trees with strong subsampling** yield optimal generalization for this dataset. The near-perfect AUROC values imply excellent separation between deceased and surviving patients, agnostic to EF. Nevertheless, accuracy and recall (especially in LightGBM and XGBoost) also remained stable, without signs of overfitting under cross-validation. The standard deviation of accuracy for LightGBM was only 0.17% across folds.

10.2. Classification Performance without Ejection Fraction

Removing EF moderately reduces discriminative power, as shown in Table 3, but performance remains clinically meaningful across nearly all models. Tree-based approaches again outperform neural and linear baselines. AUROC values remained extremely strong:

- **LightGBM AUROC:** 0.9988
- **Random Forest AUROC:** 0.9990

Notably, even without EF, **tree ensembles maintain near-perfect separability**, primarily due to complementary predictive signals from serum creatinine, age, sodium, anemia, and blood pressure. Again, the accuracy had a standard deviation of only 0.16% between folds, which indicates a lack of overfitting. As seen in Table 3, the overall rankings for the top four models remains constant, showing that some architectures have a marked advantage in with this type of data. Specifically, the consistent performance of tree-based models over the ANN classifiers suggest some level of noisy data with more "jagged" patterns, as well as reflecting the small, structured, tabular nature of the dataset itself.

Table 3. Classification Metrics without EF

Model	Acc.	Prec.	Rec.	F1
LightGBM	**98.82%**	**99.03%**	97.19%	**98.10%**
XGBoost	98.66%	98.83%	96.87%	97.84%
Random Forest	98.60%	99.09%	96.43%	97.74%
ANN Classifier	96.78%	94.94%	94.66%	94.80%
Decision Tree	97.46%	96.11%	95.79%	95.94%
KNN (k=5)	97.42%	97.19%	94.51%	95.83%
SVM (RBF)	94.20%	92.77%	88.39%	90.52%

The drop in recall seen between Tables 1 and 3 across models highlights the class imbalance. However, the AUROC measurements show that the models still successfully discriminate between patient outcomes with extremely high confidence.

10.3. Regression Performance: Ejection Fraction Prediction

The ANN regressor achieved strong performance, producing clinically useful predictions within ± 5 EF points for most patients, as illustrated in Table 4 and Figure 3.

Table 4. Ejection Fraction Regression Performance

Model	MAE	Huber Loss
ANN Regressor	**2.80**	**2.41**

The model's MAE of 2.8 significantly outperforms typical clinical estimation variability, suggesting utility as a screening aid for early risk detection and assessment. The model follows the general trend

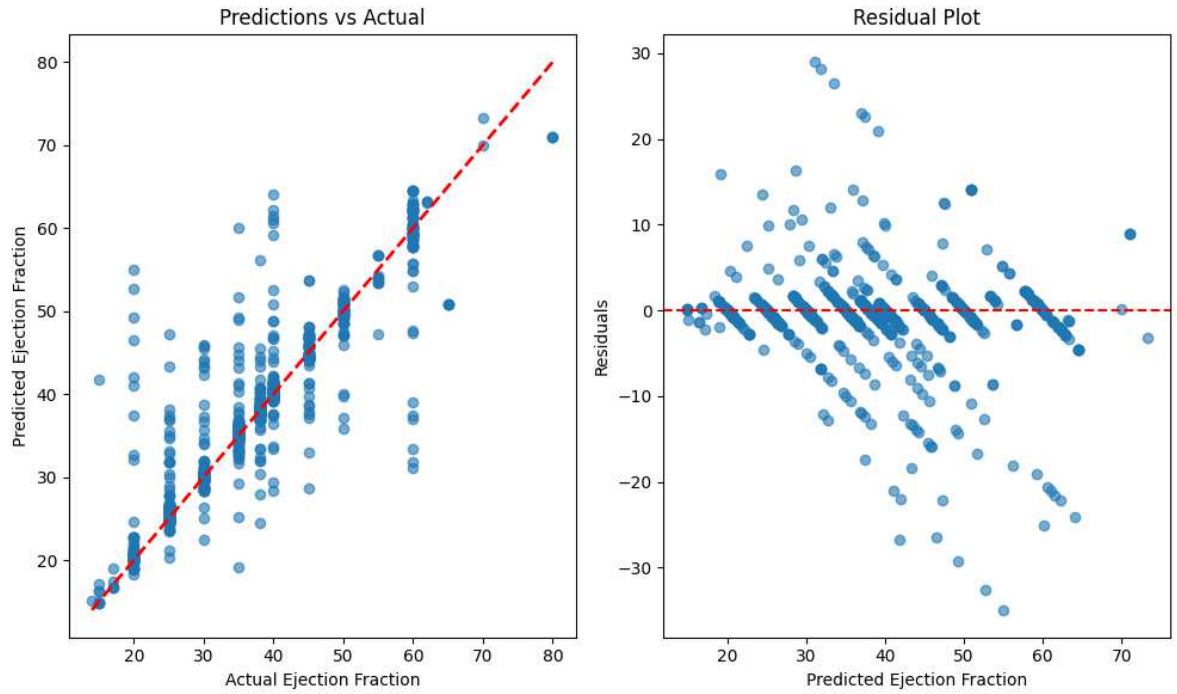

Figure 3. Plots of the Regression ANN's predictions against the actual value

of EF, as shown in Figure 3, with some discrepancies that worsen closer to the extremes. Overall, the ANN Regressor shows strong clinical utility as an aid alongside the classifier models.

10.4. Performance of Survival Analysis

The Cox Proportional Hazards model yields:

- **C-index with EF:** 75.45%

- **C-index without EF:** 72.60%

Kaplan–Meier curves generated using CoxPH stratify patients into distinct high-risk and low-risk clusters, seen in Figure 4. Although not perfectly accurate in terms of probabilities, these curves can be clinically useful for urgency analysis [48,51].

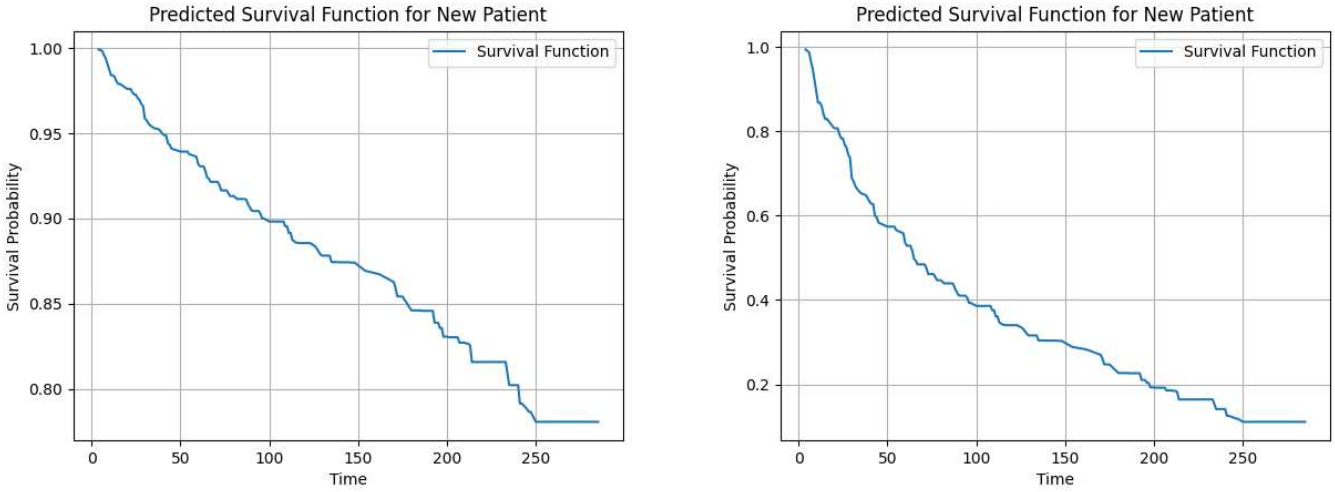

Figure 4. Risk and urgency indications by CoxPH for a relatively healthy (left) and unhealthy (right) patient

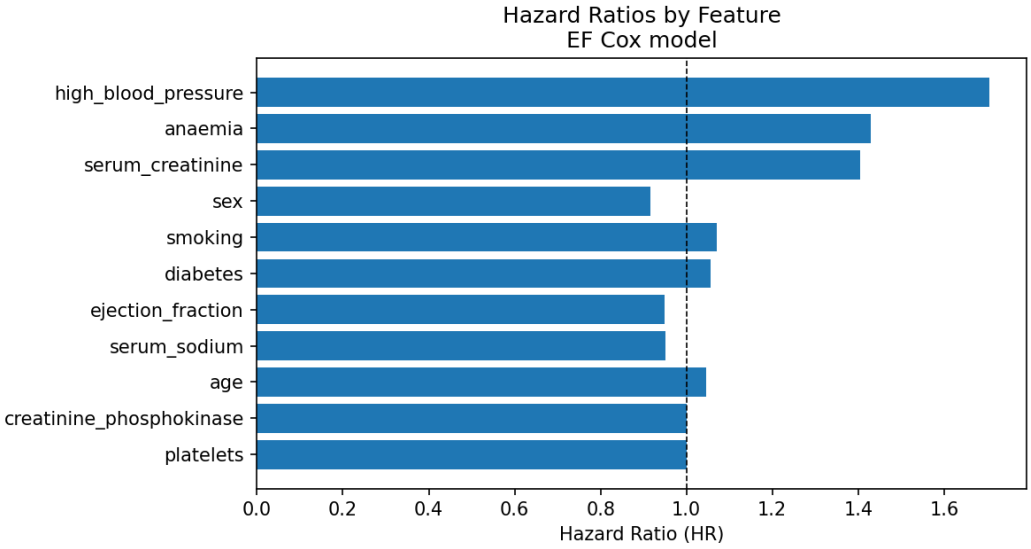

Figure 5. Hazard Ratios including Ejection Fraction

The hazard ratios, shown in Figures 5 and 6, can provide some insights into the inner workings of the cox model. Although it is not the most accurate model, this data will provide information on what features affect a patient's immediate risk of fatal HF.

10.5. Unsupervised and Multi-Output Baselines

Two additional experiments highlight the limits of unsupervised or naive model structures:

- **K-Means clustering:** 53.0% accuracy with EF, 52.9% without EF.
- **Multi-output ANN:** 80.60% classification accuracy, 3.289 MAE on EF regression.

These results confirm that mortality prediction and EF estimation require distinct, specialized model architectures.

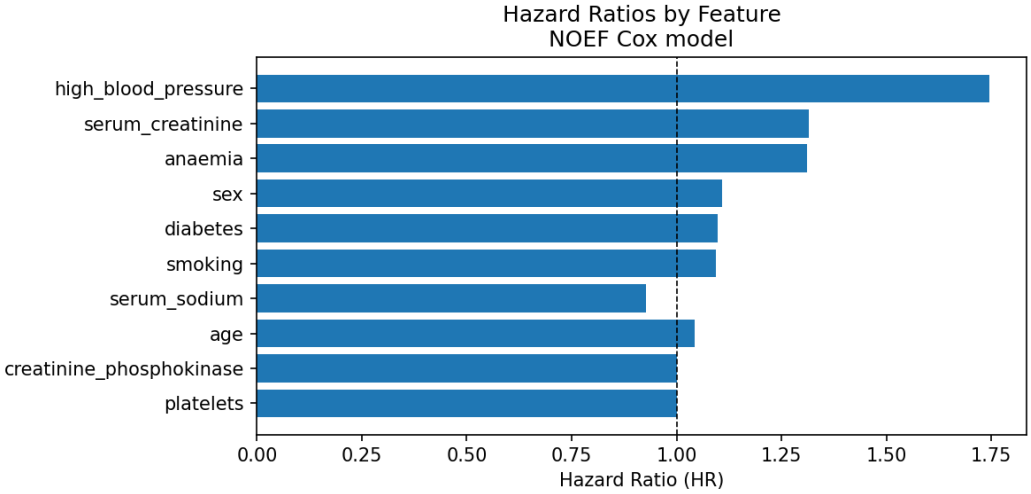

Figure 6. Hazard Ratios excluding Ejection Fraction

10.6. Summary of Key Findings

- Random Forest is the strongest classifier overall, achieving up to **98.88%** accuracy.
- ANN classifiers remain competitive, especially with EF included.
- ANN regression demonstrates clinically meaningful EF prediction (MAE 2.8).
- CoxPH offers interpretable long-horizon risk estimation, though with lower predictive power.

Overall, the combination of ensemble learning, neural models, and survival analysis provides a comprehensive multitask framework for heart failure prediction.

11. Implementation: Understanding Tree-Based Models Biologically

Across all ensemble models (RF, XGBoost, LightGBM), feature-importance analyses using SHAP, permutation AUC-drop, and impurity-based metrics consistently identified **Ejection Fraction (EF)**, **serum creatinine**, **age**, **serum sodium**, and **platelets** as the strongest predictors of heart-failure mortality. This agreement across three distinct tree-based architectures indicates that the learned representations are stable, not model-specific artifacts. Permutation AUC-drop analysis further validated this consistency: perturbing EF or serum creatinine resulted in large AUROC degradation, whereas low-ranked clinical covariates (e.g., smoking, diabetes, sex) produced minimal impact. This section interprets the biological drivers of these tree-model predictions and outlines the framework for Genetic Algorithm (GA) feature selection, which will be used to identify redundancies and dependencies among predictors.

11.1. Feature Importance

The Figures 7 and 8 show LightGBM feature-importance rankings under both EF-included and EF-excluded settings. These are supported by nearly identical rankings produced independently by XGBoost and Random Forest. All of these were calculated with AUC-drop permutation analysis [53]. **Serum Creatinine** remains the strongest predictor across all tree-based models and the ANN, reflecting, from a technical standpoint, that the models are using similar patterns for their predictions. This indicates a lack of overfitting from the Gradient Boosting trees, which are prone to this, reinforced by RF showing almost identical patterns. Between the above importances, and across other models as well, the lack of information regarding the Ejection Fraction causes heavy reliance on Serum Creatinine, as well as some reliance on Age and Serum Sodium. All other features are much less important. **Serum Creatinine** remains the strongest predictor across all tree-based models and the ANN, reflecting, from a technical standpoint, that the models are using similar patterns for their predictions. This indicates a lack of overfitting from the Gradient Boosting trees, which are prone to this, reinforced by RF showing

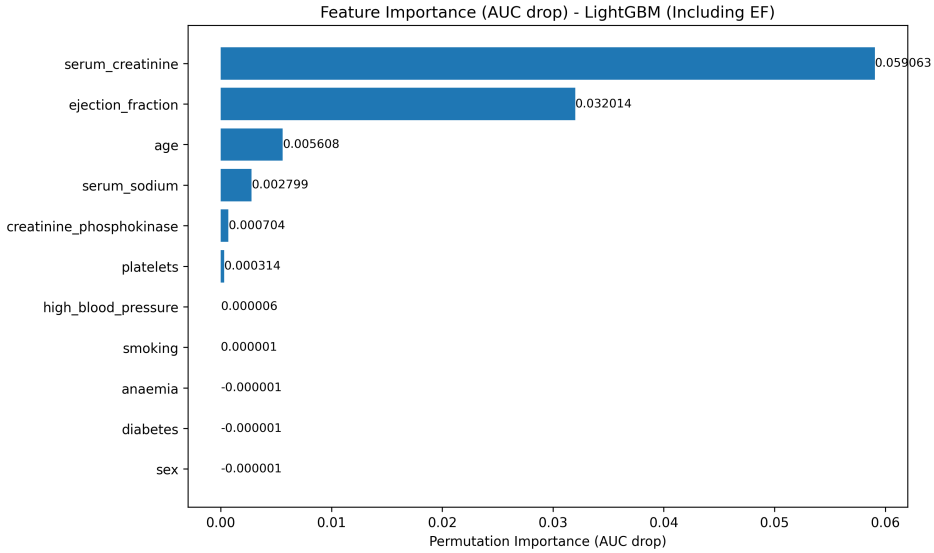

Figure 7. Permutation AUC-drop feature importance with EF included (LightGBM).

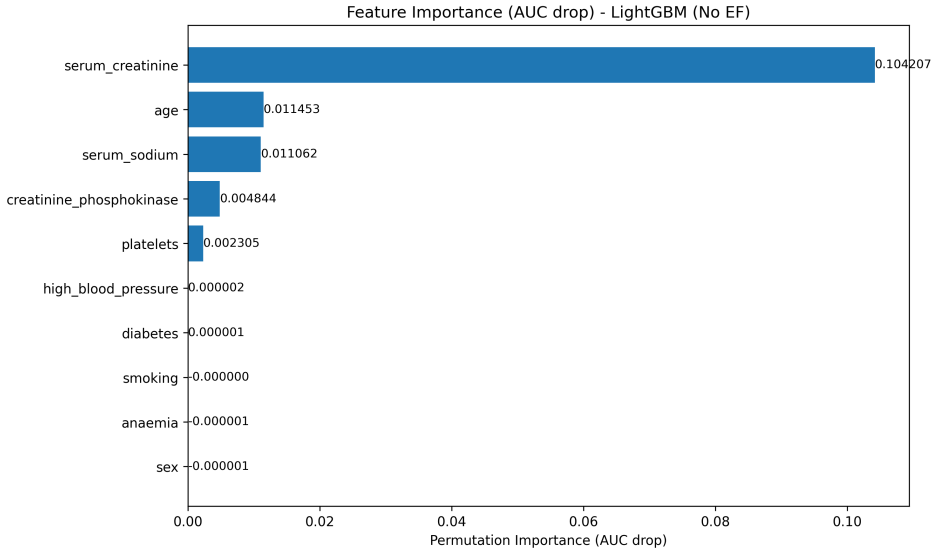

Figure 8. Permutation AUC-drop feature importance without EF (LightGBM).

almost identical patterns. Between the above importances, and across other models as well, the lack of information regarding the Ejection Fraction causes heavy reliance on Serum Creatinine, as well as some reliance on Age and Serum Sodium. All other features are much less important.

11.2. Identifying Redundant or Dependent Features Using Genetic Algorithm

To assess which features are essential and which may be redundant, a Genetic Algorithm (GA)–based feature-selection framework was developed. This framework is applicable to any model, but it will be run on LightGBM, our best model, and Random Forest, which is non-gradient boosted and so less prone to overfitting.

11.2.1. Implementation Framework

Each chromosome represents a binary feature-selection mask across ten clinical features (excluding time, death event, and EF). For each chromosome, a new classifier is trained using only the selected subset. The pipeline is:

- Initial population of 100 feature subsets (including a full-feature baseline).
- Top 10% of each generation propagate unchanged (elitism).

- 1000 mutated offspring generated via probability-weighted selection.
- Convergence criterion: identical best chromosome for 5 consecutive generations.
- All evaluated subsets cached to avoid repeated training.

This setup ensures a balance between exploration of the feature space and exploitation of promising subsets, while maintaining computational efficiency.

11.2.2. Results using LightGBM

After running the algorithm, it took 12 iterations to converge with LightGBM on the following features: **Serum Creatinine, Serum Sodium, Age, and High Blood Pressure**. Upon further analysis, every model in the top ten of the final survival pool utilized these features, seen clearly in Figure 9. To test if this a product of overfitting, the same was run with Random Forest, a model much less susceptible to overfitting than gradient-boosted trees, for near identical results. It, too, took 12 iterations to converge on the same four features, which all of its top ten used as well. This is unsurprising given the near-identical feature importance rankings amongst the tree-based models.

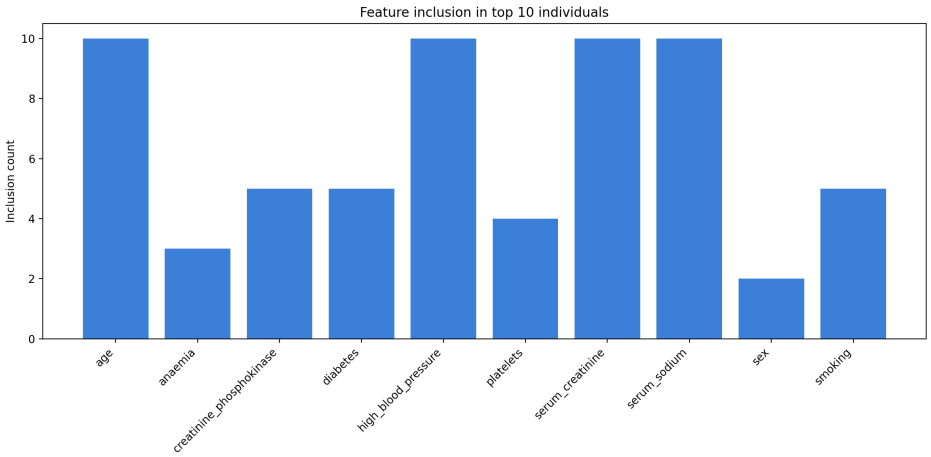

Figure 9. Feature usage frequency across the top ten GA-surviving chromosomes (LightGBM).

These results echo a strong conclusion that these four features are strongly linked to the risk of Heart Failure. For three of the four, the result can be explained trivially from a technical standpoint: Serum Creatinine (SC), Serum Sodium (SS), and Age were the three most important predictors in the absence of EF, shown in Figure 8, and other features may have acted as noise. High Blood Pressure (HBP), however, is harder to explain. The winning model beat the initial no-EF model's accuracy of 98.82% with **99.04%**, although its AUROC did drop from 0.9988 to **0.9958**. To verify that these four features were indeed better than just three, a LightGBM model was trained using only SC, SS, and Age. It performed somewhat well, with a 97.76% accuracy and a 0.9934 AUROC, but HBP clearly did provide the winning model with useful data. Although not present at the top of the feature importance rankings for the tree-based models such as LightGBM, HBP was the greatest Hazard Ratio on the CoxPH model, shown with and without EF in Figures 5 and 6 respectively. Therefore, CoxPH's accurate classification of high- and low-risk patients strongly relied on HBP, which could explain its significance in the LightGBM tree-based model despite its absence from a top spot in the feature importance ranking.

11.3. Clinical Significance of the Dominant Features when excluding EF

Across all tree-based models, and even the ANN predictor, Serum Creatinine reigned as the most valuable predictor. In fact, for every 0.5mg/dl rise in Creatinine concentration over the average, the mortality probability rises by 15% on average [54]. These statistics justify the importance placed on Creatinine, agnostic to the presence of Ejection Fraction. These statistics also reflect the drastic rise in Average Serum Creatinine shown between patients who survived and those who passed away

in Figure 1 for reference. Biologically, this happens because during HF, reduced cardiac output and congestive venous pressures impair renal perfusion. This leads to renal hypoperfusion, ischemia, and kidney injury, which reflect in elevated Creatinine levels [55].

When Ejection Fraction is not present, Age and Serum Sodium also increase in importance. Aging causes the cardiac structure to degrade, which non-trivially causes the risk of HF to rise [56]. This aligns with the rise in average age between patients who survived, and those who did not, in Figure 1. Serum Sodium, especially when reduced below normal levels at 137.5 mmol/L (hyponatremia), is also seen as an indicator of HF, since for every 1 mmol/L decrease below average, a 4% rise in mortality risk occurs, reflecting our dataset in Figure 1 once again [57]. This is because during HF, reduced cardiac output and effective circulatory volume, the body releases non-osmotic Arginine Vasopressin, causing greater water retention with no change in Sodium, leading to decreased Serum Sodium concentrations, a strong indication of HF [58].

Technically speaking, High Blood Pressure was always the biggest outlier, due to the fact that it did not occupy a top spot in the feature importance rankings. However, our CoxPH model did identify it as an important feature, hence its selection in the Genetic Algorithm test (GA). Biologically, chronic HBP is a direct cause of HF, causing structural changes in the left ventricle of the heart, since the pressure overload causes the myocardium to increase the thickness of the arterial wall (hypertension) [59].

12. Web application with the Integration of Gemini

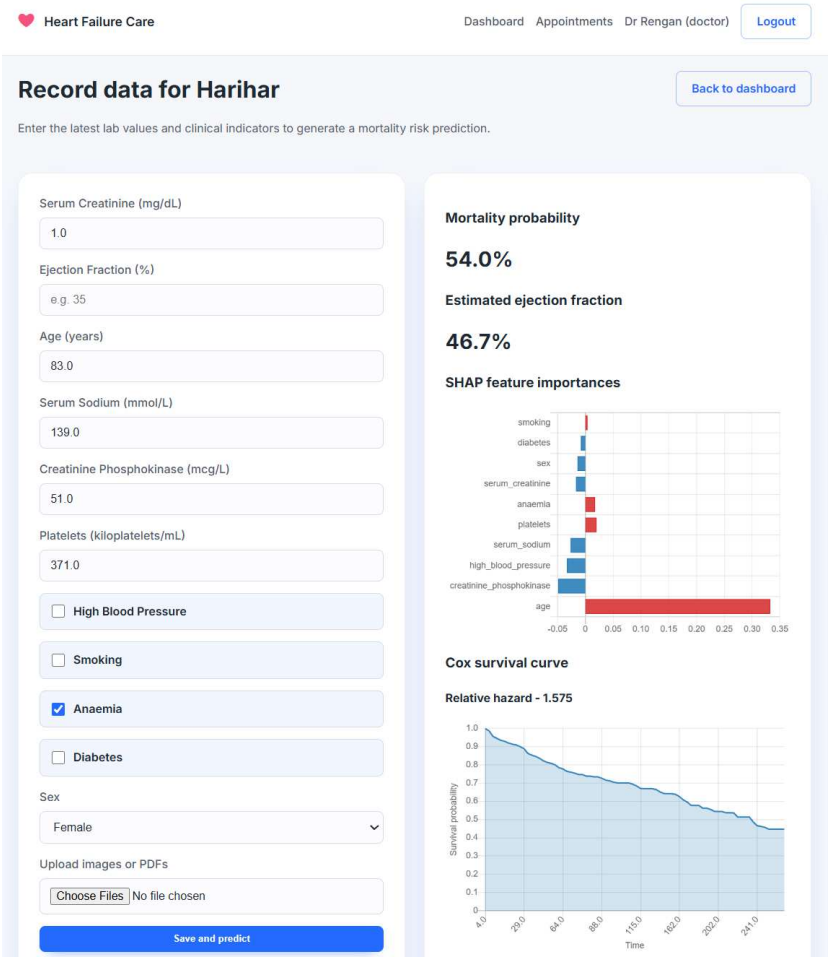

Figure 10. The patient input dashboard, our core feature

We built a web application interface around this app to make clinical use easy and maximize impact. As seen in Figure 10, doctors will be able to input details about their patient or upload a PDF / image file. If a file is uploaded, a Gemini-based OCR built with n8n reads the needed information and

completes the input fields.

Once the information is taken in, our models make the needed predictions. Firstly, a mortality probability is predicted alongside a SHAP feature importance graph, where larger red features had a greater impact on increasing mortality, helping doctors understand the model's reasoning and make logical suggestions to their patients. If the patient does not provide Ejection Fraction, this of course affects which model is used, with a minute change in accuracy nonetheless, as seen in Tables 1 and 3. When not provided, our model is able to predict their ejection fraction with an MAE of 2.80 as seen in Table 4. After this, a cox survival hazards plot is also shown, allowing doctors to understand the urgency and need for immediate treatment on every case. This model assists doctors with that and also adapts to the presence of Ejection Fraction. The C-index drops marginally here as well, however. We also use an LLM to provide further clarification on our models decisions and reasoning. We pass the inference and SHAP results to a Gemini API-based LLM using n8n for further reasoning and clinical support. Around this core feature, we have built basic signup and login systems for doctors and patients, as well as appointment booking functionality. Our aim is to ensure that our models are easy to use and harness to better the lifes of and reduce the risks for patients. Our interface utilizes a simple HTML/CSS frontend with JS, as well as a Flask Python backend. The models were trained via a Python kernel in Jupyter Notebook and exported, after which the savefiles are read for weights during inference. Additional features such as OCRs and LLMs utilize Gemini APIs, effectively integrated with an n8n system.

12.1. Hart Failure System Summary

By eliminating less important features in a Genetic Algorithm alongside AUC-drop permutation analysis for Feature Importance and CoxPH-native Hazard Ratios, we were able to arrive at the most influential features for the model's performance, namely Serum Creatinine, Age, Serum Sodium, and High Blood Pressure. The strong biological reasoning for the selection of these features, alongside their appearance under rigorous technical processes reflect how well successful models, such as Tree-based Classifiers like LightGBM or ANN Classifiers generalized to the data and extracted the underlying complex and non-linear patterns in the data for highly accurate non-invasive diagnoses of HF. The final LightGBM model was able to predict HF with an accuracy of 99.04% without the need of an Echocardiogram to determine EF.

We would also discuss about the limitations of our work. The dataset is small and gathered from a specific place which can lead to possible biases and limit the generalizability of our findings. Second, the number of features is limited. Important clinical factors such as previous heart conditions were not included. ECG results could have allowed us to spot heart failure related abnormalities, and tracking the development of blood test results over time could have allowed us to ensure that values extracted are actually meaningful. Future work would need to incorporate those features alongside other relevant ones into our model and enlarge the dataset to include a more diverse and balanced population, so that the model's early heart failure detection can become a clinically useful risk assessment and more hence reliable . Testing again over a set of real patients that were not part of our dataset would then be crucial to see if the results would match their relative echocardiographic results, giving us a more accurate picture of how much our model is really able to generalize.

13. Technical Discussion

Our findings illustrate several key observations:

- Tree-based algorithms, especially gradient boosting such as LightGBM, performed best on this dataset, with similar performance after removing noisy features.
- Neural networks approach but do not exceed ensemble performance, including Random Forest, in this dataset, which reflects the small, tabular nature of the data as well as the "jagged" class separation.

- Serum Creatinine was the strongest predictor for the tree-based models and the ANN Classifiers, agnostic to the presence of Ejection Fraction
- Survival modeling captures longitudinal mortality trends not visible in classification alone, and uses High Blood Pressure as it the strongest Hazard Ratio.
- Serum Creatinine, Serum Sodium, Age, and High Blood Pressure were sufficient to obtain a peak in accuracy for the model, reflecting the biological importance of these predictors.

14. Conclusion

This work explores multiple machine learning and statistical approaches for heart failure prediction using a unified dataset and evaluation framework. The results demonstrate that ensemble methods specifically Random Forests outperform neural networks in small clinical tabular datasets [44]. The features isolated by the genetic model, mostly in agreement with feature importance and hazard ration, reflected the modern biological consensus for factors affecting heart failure.

Author Contributions: Vinod Pangracious contributed to research study, problem formulation, workflow development 1 and 4. Harihar Rengan contributed to the development of design methodologies, model training and web application design and development 7 and 9.6, George Abdallah and Hyunmin Park contributed the design and development machine learning algorithms and results analysis 7, 9.6

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review/evaluation Board of School of Engineering, American University of Dubai. This project was completed as an undergraduate capstone project with the authorization from American University in Dubai under the approval number EC02. Since we used an open-source database and data augmentation, no human subjects were involved in the study

Informed Consent Statement: Patient consent was waived due to an open-source dataset and data augmentation

Data Availability Statement: The dataset used for this study is attached with the manuscript as suggested Data Availability Statements are available in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

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

Abbreviations

The following abbreviations are used in this manuscript:

HF	Heart Failure
CPK	Creatinine Phosphokinase
MLP	Multilayer Perceptron
XGBoost	Extreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
RBF	Radial Basis Function
GA	Genetic Algorithm
SHAP	SHapley Additive exPlanations
PH	Proportional Hazards
CoxPH	Cox Proportional Hazards Model
AUROC	Area Under the Receiver Operating Characteristic Curve
AUPRC	Area Under the Precision–Recall Curve
ECE	Expected Calibration Error
MOTE	Synthetic Minority Over-sampling Technique
HDF5	Hierarchical Data Format version 5
WCAG	Web Content Accessibility Guidelines
HIPAA	Health Insurance Portability and Accountability Act
GDPR	General Data Protection Regulation

References

1. National Health Service (NHS) UK, Heart failure Overview and Causes, Online: https://www.nhs.uk/conditions/heart-failure/, Page last reviewed 19 May 2022, accessed 10 Oct 2025
2. Saqib M, Perswani P, Muneem A, Mumtaz H, Neha F, Ali S, Tabassum S. Machine learning in heart failure diagnosis, prediction, and prognosis: review. Ann Med Surg (Lond). 2024 May 6;86(6):3615-3623. doi: 10.1097/MS9.0000000000002138. PMID: 38846887; PMCID: PMC11152866.
3. S. W. Siddiqui, T. Ashok, N. Patni, M. Fatima, A. Lamis, and K. K. Anne, "Anemia and heart failure: A narrative review," *Cureus*, vol. 14, no. 7, p. e27167, 2022.
4. Oktay AA, Paul TK, Koch CA, et al. Diabetes, Cardiomyopathy, and Heart Failure. [Updated 2023 Sep 26]. In: Feingold KR, Adler RA, Ahmed SF, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2000-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK560257/
5. Ziegler M, Wang X, Peter K. Platelets in cardiac ischaemia/reperfusion injury: a promising therapeutic target. Cardiovasc Res. 2019 Jun 1;115(7):1178-1188. doi: 10.1093/cvr/cvz070. PMID: 30906948; PMCID: PMC6529900.
6. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. JAMA Netw Open. 2020 Jan 3;3(1):e1918962. doi: 10.1001/jamanetworkopen.2019.18962. PMID: 31922560; PMCID: PMC6991258.
7. Ding N, Shah AM, Blaha MJ, Chang PP, Rosamond WD, Matsushita K. Cigarette Smoking, Cessation, and Risk of Heart Failure With Preserved and Reduced Ejection Fraction. J Am Coll Cardiol. 2022 Jun 14;79(23):2298-2305. doi: 10.1016/j.jacc.2022.03.377. PMID: 35680180; PMCID: PMC12711428
8. Heart Failure Clinical Records [Dataset]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5Z89R.
9. E. Glazier and E. Ko, Smokers twice as likely to develop congestive heart failure, UCLA Health-Ask the Doctors, News Article, Accessed 10 Oct 2025, Apr. 2024.
10. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak. 2020 Feb 3;20(1):16. doi: 10.1186/s12911-020-1023-5. PMID: 32013925; PMCID: PMC6998201.
11. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70:1321–1330, 2017.
12. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019 Dec 16;17(1):230. doi: 10.1186/s12916-019-1466-7. PMID: 31842878; PMCID: PMC6912996.
13. Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, Parker Barnes, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in Proceedings of the 2020 Conference onFairness, Accountability, and Transparency (FAT*), 2020, pp. 33–44
14. Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
15. Tianqi Chen, Carlos Guestrin "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794
16. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995, https://doi.org/10.1007/BF00994018.
17. B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2002, pp. 694–699.
18. I. Mahmud, M. M. Kabir, M. F. Mridha, S. Alfarhood, M. Safran, and D. Che, "Cardiac failure forecasting based on clinical data using a lightweight machine learning metamodel," Diagnostics (Basel), vol. 13, no. 15, p. 2540, Jul. 2023.
19. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, e0118432, 2015.
20. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006 Nov-Dec;26(6):565-74. doi: 10.1177/0272989X06295361. PMID: 17099194; PMCID: PMC2577036.
21. S. M. Lundberg et al., "From local explanations to global understanding with explainable ai for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.

22. Amershi, Saleema and Weld, Dan and Vorvoreanu, Mihaela and Fourney, Adam and Nushi, Besmira and Collisson, Penny and Suh, Jina and Iqbal, Shamsi and Bennett, Paul N. and Inkpen, Kori and Teevan, Jaime and Kikin-Gil, Ruth and Horvitz, Eric, "Guidelines for human-ai interaction," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–13.
23. E. W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd ed. Cham: Springer, 2019, ebook.
24. D. E. Austin et al., "Comparison of machine learning and the regression-based EHMRG model for predicting early mortality in acute heart failure," International Journal of Cardiology, vol. 365, pp. 78–84, 2022.
25. A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (KDD), 2005, pp. 625–632.
26. H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239
27. Wu, L, Rizwan, A, Rodriguez, M. et al. Heart Failure With Mildly Reduced Ejection Fraction: A Call for a Precision Medicine Approach. JACC Adv. 2026 Jan, 5 (1) . https://doi.org/10.1016/j.jacadv.2025.102476
28. Hajishah, H., Kazemi, D., Safaee, E. et al. Evaluation of machine learning methods for prediction of heart failure mortality and readmission: meta-analysis. BMC Cardiovasc Disord 25, 264 (2025). https://doi.org/10.1186/s12872-025-04700-0
29. J.-H. Kim, P. Shah, U. S. Tantry, and P. A. Gurbel, "Coagulation abnormalities in heart failure: Pathophysiology and therapeutic implications," *Current Heart Failure Reports*, vol. 13, no. 6, pp. 319–328, 2016.
30. G. Savarese and L. H. Lund, "Global burden of heart failure: A comprehensive and updated review of epidemiology," *Cardiovascular Research*, vol. 118, no. 17, pp. 3272–3287, 2022.
31. P. Ponikowski et al., "2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure," *European Heart Journal*, vol. 37, no. 27, pp. 2129–2200, 2016.
32. P. A. Heidenreich et al., "2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure,"*Journal of the American College of Cardiology*, 2022.
33. G. Savarese and L. H. Lund; P. Ponikowski et al., "Chronic HF natural history discussions," *Cardiovascular Research*; ESC, 2016 https://doi.org/10.1093/eurheartj/ehw128.
34. P. A. Heidenreich et al., "Heart failure etiologies and definitions,"AHA/ACC/HFSA Guidelines, 2022.
35. P. A. Heidenreich et al., "Complications and multiorgan congestion in HF," AHA/ACC/HFSA, 2022.
36. Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452
37. P. A. Heidenreich et al., Guideline for the Management of Heart Failure: Executive Summary: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines, 2022, P e876-e894, V 145, N 18, AHA/ACC/HFSA, 2022.
38. T. D. Filippatos et al., "Hyponatremia in patients with heart failure," *European Journal of Heart Failure*, 2013.
39. Mojadidi MK, Galeas JN, Goodman-Meza D, Eshtehardi P, Msaouel P, Kelesidis I, Zaman MO, Winoker JS, Roberts SC, Christia P, Zolty R. Thrombocytopaenia as a Prognostic Indicator in Heart Failure with Reduced Ejection Fraction. Heart Lung Circ. 2016 Jun;25(6):568-75. doi: 10.1016/j.hlc.2015.11.010. Epub 2016 Jan 18. PMID: 26868831.
40. Stefane Souza, V., & Araújo Lima, D. (2024). Cardiac Disease Diagnosis Using K-Nearest Neighbor Algorithm: A Study on Heart Failure Clinical Records Dataset. Artificial Intelligence and Applications, 3(1), 56-71. https://doi.org/10.47852/bonviewAIA42022045
41. Qadri AM, Hashmi MSA, Raza A, Zaidi SAJ, Rehman Au. 2024. Heart failure survival prediction using novel transfer learning based probabilistic features. PeerJ Computer Science 10:e1894 https://doi.org/10.7717/peerj-cs.1894
42. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019 Apr 4;380(14):1347-1358. doi: 10.1056/NEJMra1814259. PMID: 30943338.
43. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019 Jan;25(1):44-56. doi: 10.1038/s41591-018-0300-7. Epub 2019 Jan 7. PMID: 30617339.
44. Leo Grinsztajn and Edouard Oyallon and Gael Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" arXiv preprint, 2022. https://arxiv.org/abs/2207.08815.
45. Ravid Shwartz-Ziv, Amitai Armon, "Tabular data: Deep learning is not all you need", Information Fusion, Volume 81, 2022, Pages 84-90, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2021.11.011.
46. Huber, P.J. (1964) Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 35, 73-101. https://doi.org/10.1214/aoms/1177703732

47. Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2 (IJCAI'95). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1137–1143.
48. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996 Feb 28;15(4):361-87. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4. PMID: 8668867.
49. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011 May 10;30(10):1105-17. doi: 10.1002/sim.4154. Epub 2011 Jan 13. PMID: 21484848; PMCID: PMC3079915.
50. D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B*, Pages 187-202, 1972, https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.
51. Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452
52. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011 May 10;30(10):1105-17. doi: 10.1002/sim.4154. Epub 2011 Jan 13. PMID: 21484848; PMCID: PMC3079915.
53. Lundberg, Scott M and Lee, Su-In, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, 2017. volume 30, 2017.
54. G. L. Smith et al., "Renal impairment and outcomes in heart failure," *Journal of the American College of Cardiology*, 2006.
55. Damman K, Testani JM. The kidney in heart failure: an update. Eur Heart J. 2015 Jun 14;36(23):1437-44. doi: 10.1093/eurheartj/ehv010. Epub 2015 Apr 2. PMID: 25838436; PMCID: PMC4465636.
56. Li H, Hastings MH, Rhee J, Trager LE, Roh JD, Rosenzweig A. Targeting Age-Related Pathways in Heart Failure. Circ Res. 2020 Feb 14;126(4):533-551. doi: 10.1161/CIRCRESAHA.119.315889. Epub 2020 Feb 13. PMID: 32078451; PMCID: PMC7041880.
57. S. Peng, J. Peng, L. Yang, and W. Ke, "Relationship between serum sodium levels and all-cause mortality in congestive heart failure patients: A retrospective cohort study based on MIMIC-III," *Frontiers in Cardiovascular Medicine*, vol. 9, 2023.
58. S. E. Ishikawa, "Hyponatremia associated with heart failure: Pathological role of vasopressin-dependent impaired water excretion," *Journal of Clinical Medicine*, vol. 4, no. 5, pp. 933–947, 2015.
59. G. C. Oh and H. J. Cho, "Blood pressure and heart failure," *Clinical Hypertension*, vol. 26, p. 1, 2020.