Journal of High School Science


# Machine learning for assessing the role of immunoglobulin free light chains in the identification of circulatory diseases and survival prediction

Munteanu B

Submitted: September 2, 2025, Revised: version 1, October 13, 2025, version 2, November 5, 2025, version 3, November 11, 2025
Accepted: December 10, 2025

**Abstract**

Recent studies have highlighted the predictive value of combined Immunoglobulin Free Light Chain (FLC) levels in circulatory disease (CD), as well as other life-threatening conditions. Elevated FLC levels have been associated with poor outcomes, including increased mortality and hospital readmissions. In acute Heart Failure (HF), for example, patients had significantly higher FLC levels compared to those with stable HF or healthy controls. Moreover, FLC levels remained elevated long after clinical stabilization, suggesting potential ongoing immune disturbances, or impaired renal clearance. The present study investigated the use of Machine Learning (ML) techniques to predict the survival of patients with elevated FLC and to improve the understanding of circulatory mortality risks by identifying individuals whose deaths were attributable to circulatory system conditions. It was found that binary classification models could successfully determine patients with negative outcome (AUC > 0.80), and survival analysis models could achieve dynamic AUC higher than 0.75 in estimating the survival probability as a function of time. In the context of CD, elevated levels of FLCs have been shown to be associated with an increased risk of circulatory events. This association suggests that FLCs might play a role in the development of circulatory conditions.

**Keywords**

Free Light Chains, Circulatory Disease, Machine Learning, Survival, Support Vector Machine, Random Forest, XGBoost classifier, XGBoost survival embeddings, SMOTE, Label Encoder

Briana Munteanu, Granada High School, 400 Wall St, Livermore, CA 94550, USA brianamunteanu1@gmail.com


https://doi.org/10.64336/001c.154326

## 1. Introduction

**1.1** Role of immunoglobulin Free Light Chains in disease identification

Circulatory disease (CD) remains the leading cause of morbidity and mortality worldwide (1). Despite advancements in medical interventions and preventive strategies, CD continues to impact millions of individuals each year, contributing to high healthcare costs and reduced quality of life. Early detection and accurate risk stratification remain critical challenges in preventing adverse circulatory outcomes. In recent years, there has been growing interest in identifying reliable biomarkers that can predict disease progression, inform treatment strategies, and assess therapeutic effectiveness. (2, 3).

Among emerging biomarkers, immunoglobulin free light chains (FLCs) have gained attention due to their potential role in circulatory pathology. Immunoglobulins (commonly known as antibodies) are composed of two types of protein chains: heavy and light chains. Each antibody contains only one type of light chain—either *kappa* or *lambda*. Although *kappa* and *lambda* light chains share similar structures and functions, they are encoded by distinct genes. In a typical antibody, light chains pair with heavy chains to form a complete structure. However, when light chains are produced in excess, some may not be incorporated into antibodies. These excess light chains, which circulate freely in the blood, are known as FLCs. A small amount of FLCs is normally present in the blood, at low concentrations, in healthy individuals (4). Abnormally elevated levels of FLCs, particularly when there is an imbalance between *kappa* and *lambda* chains, are commonly associated with conditions such as multiple myeloma, Monoclonal Gammopathy of Undetermined Significance (MGUS), and light-chain (AL) Amyloidosis (5-10). Moreover, elevated FLC levels are associated with a broad range of circulatory conditions, including HF, coronary artery disease, and systemic inflammation (11-16).

Chronic inflammation is a well-established driver of atherosclerosis, endothelial dysfunction, and myocardial injury (17). FLCs are implicated in inflammatory cardiovascular processes, with studies demonstrating a positive correlation between elevated FLC levels and inflammatory biomarkers such as C-Reactive Protein (CRP), a marker of systemic inflammation and cardiovascular risk (18), and Interleukin-6 (IL-6), a cytokine involved in vascular inflammation and endothelial damage (19).

In addition, elevated *kappa* and *lambda* FLC levels are linked to increased cardiovascular mortality, possibly due to their role in inflammatory signaling and endothelial dysfunction (20, 21). FLCs also promote oxidative stress and vascular remodeling which can contribute to hypertension and arterial stiffness (12).

Elevated FLC levels are commonly observed in patients with HF with preserved Ejection Fraction (HFpEF) and HF with reduced Ejection Fraction (HFrEF) (20). These increased FLC levels are linked to ventricular dysfunction (12) and myocardial fibrosis, as FLCs may contribute to fibrotic remodeling that impairs cardiac function over time (19). Additionally, some studies suggest that FLC levels can predict

hospitalization and mortality in HF patients (18).

In light-chain Amyloidosis (AL), misfolded FLCs aggregate and deposit in the myocardium, leading to restrictive cardiomyopathy and progressive HF (6, 22). FLC elevation has also been linked to increased risk of arrhythmias, including atrial fibrillation (18), ventricular tachycardia and fibrillation (20). Some patients with high FLC levels develop conduction blocks, which may require pacemaker implantation (6). Given these associations, FLCs could serve as predictive biomarkers for arrhythmic events and sudden cardiac death, particularly in high-risk populations (23).

Despite significant progress in the diagnosis and treatment of CD over the past several decades, many challenges remain. This study seeks to leverage the advantages of algorithmic approaches to investigate and enhance the use of FLCs measured levels in distinguishing CD patients. While FLCs show great promise as biomarkers for CD, further research is needed to determine the most effective ways to use them in distinguishing between high-risk individuals and those with early-stage or advanced CD.

**1.2** Role of Machine Learning algorithms in disease progression prediction

The increasing interest in applying machine learning (ML) to biomarker-driven diagnostics has raised valid questions regarding its clinical utility, particularly when weighed against the performance and interpretability of established diagnostic cutoffs. Clinicians may be justifiably skeptical of adopting a "black box" algorithm whose predictions offer no mechanistic transparency and whose marginal gains in predictive resolution do not affect downstream therapeutic decisions (24). However, one key strength of ML models is their ability to model complex, non-linear, and continuous relationships without relying on hard cutoffs like traditional rule-based systems (25). This may help avoid situations where a patient just below a cutoff is wrongly classified as "low risk" despite having a worrisome overall profile. Statistical methods that rely on separating patients in groups and then determining a common prognosis model for each group are dependent on the subjectivity of groups' definition and do not provide the means for individual characterization.

For example, quartile-based discretization is a common statistical practice (26, 27), but it inherently flattens the data by reducing rich, continuous variation into broad categorical bins. This simplification results in a loss of granularity, as meaningful differences within each quartile are ignored, treating all values within a range as equivalent. Such an approach can obscure subtle but important trends, weaken associations, and distort relationships between variables - especially in skewed distributions where quartiles may not represent evenly spaced or meaningful value ranges. In contrast, ML algorithms can handle continuous biomarker data directly, preserving the full variability and structure of the data without imposing arbitrary thresholds or groupings. This allows them to capture complex, nonlinear relationships and subtle patterns that might be lost through quartile-based categorization. Additionally, many ML methods are robust to skewed distributions and can learn from the data as is, making them more flexible and informative in scenarios where classical statistical approaches

may oversimplify or distort the underlying signals (25).

Thus, this study used ML techniques to explore the predictive value of FLCs in identifying patients at risk of CD- related mortality. The specific objectives were as follows: [1] to construct a binary classification model that differentiates patients based on survival status (*dead* vs. *right-censored* patients) within a 5,000-day timeframe; [2] to improve the understanding of circulatory mortality risks by identifying individuals whose deaths were attributable to circulatory system conditions, as opposed to non-circulatory causes; and [3] to estimate the survival probability over time for those patients with elevated FLC blood concentration.

Recent work by Ren et al. (28) explored the application of shallow ML techniques, demonstrating that trained models could effectively distinguish between CD patients and healthy individuals. The present study differs from this prior work in three key aspects: [1] it utilizes a smaller set of input features for model training (6 vs. 11 features), [2] it provides evidence that the ML models can also differentiate between patients with elevated FLC values who experienced negative outcomes —regardless of the cause of death—and those who survived, and [3] it demonstrates the feasibility of accurately estimating survival probability over time in patients identified as being at risk.

## 2. Materials and Methods

### **2.1** Data description

The dataset used in this study was a subset of publicly available data originally collected by (12) between January 1, 1995, and November 21, 2003, and it was included in the *scikit-survival* Python module. It consisted of 7,874 patients and included nine variables: *age*, *chapter*, *creatinine*, *flc.grp*, *kappa*, *lambda*, *mgus*, *sample.yr*, and *sex*. The *chapter* variable categorized the cause of death into 16 distinct groups (Circulatory=745, Neoplasms= 567, Respiratory= 245, Mental= 144, Nervous= 130, Digestive= 66, External Causes= 66, Endocrine= 48, Genitourinary= 42, Ill Defined= 38, Infectious= 32, Injury and Poisoning= 21, Musculoskeletal= 14, Blood= 4, Skin= 4, and Congenital=3). Among all patients, 2,169 experienced the event (death), whereas 5,705 were right censored. Additionally, the dataset included the observation time (in days), and a binary event indicator denoting whether the patient died or if the observation was right censored. Censoring is a fundamental aspect of survival analysis, referring to situations where the exact time of an event (i.e., death, failure, or relapse) is not observed. The most common type and the one relevant to our case was right censoring which occurs when the event has not happened by the end of the study or when a subject leaves the study early.

The histogram in Figure 1 visualizes the distribution of observation times for both groups, providing insights into survival patterns within the cohort.

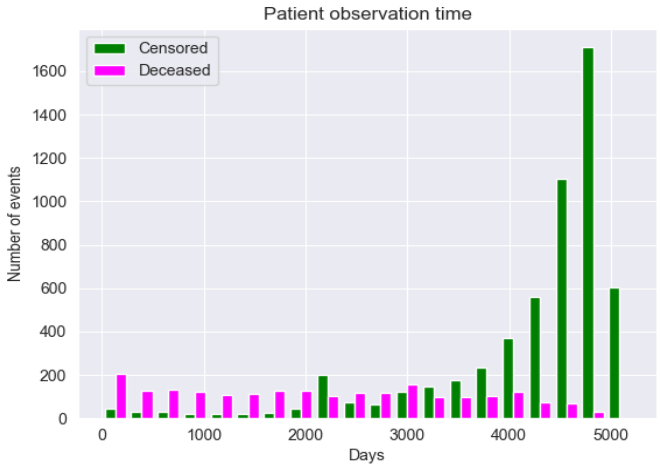


**Figure1**. Distribution of deceased and right-censored patients

The recorded time (either until death, or right censored) ranged from 0 to 4,998 days, with circulatory conditions being the most common cause (745 cases). The variable *flc.grp* represented the original classification of patients into 10 categories based on the values of the FLC components *kappa* and *lambda* (12). According to (12), patients with an elevated sum of kappa and lambda chains have a demonstrably worse overall survival chance than patients without plasma cell disorders. In the original analysis, the participants were divided into 10 distinct groups based on their total FLC blood concentration. Survival curves for each group showed a progressive reduction in survival probability, with those in group 10 having the fastest decrease.

The dataset included 1,350 patients with missing creatinine levels. Among the nine measured variables, six (*age*, *creatinine*, *kappa*, *lambda*, *flc.grp*, and *sample.yr*) were continuous, while three (*chapter*, *mgus,* and *sex*,) were categorical. At baseline, the median age was 63 years (range, 50-101 years), and 55.24 % (n= 4350) were women. Descriptive statistics were computed in Table 1 and Table 2.

Python programming language version 3.12.3 was used for data processing, visualization, and model building.

**Table 1.** Descriptive statistics for categorical variables

| Variable | Category | Frequency (n) | Percentage (%) |
|---|---|---|---|
| sex | male | 3524 | 44.75% |
| | female | 4350 | 55.25% |
| chapter | circulatory | 745 | 34.35% |
| | neoplasms | 567 | 26.14% |
| | respiratory | 245 | 11.30% |
| | mental | 144 | 6.64% |
| | nervous | 130 | 5.99% |
| | external causes | 66 | 3.04% |
| | digestive | 66 | 3.04% |
| | endocrine | 48 | 2.21% |
| | genitourinary | 42 | 1.94% |
| | ill defined | 38 | 1.75% |
| | infectious | 32 | 1.48% |
| | injury and poisoning | 21 | 0.97% |
| | musculoskeletal | 14 | 0.65% |
| | blood | 4 | 0.18% |
| | skin | 4 | 0.18% |
| | congenital | 3 | 0.14% |
| flc.grp | 1st decile | 679 | 8.75% |
| | 2nd decile | 798 | 10.28% |
| | 3rd decile | 816 | 10.52% |
| | 4th decile | 783 | 10.09% |
| | 5th decile | 787 | 10.14% |
| | 6th decile | 790 | 10.18% |
| | 7th decile | 806 | 10.39% |
| | 8th decile | 730 | 9.41% |
| | 9th decile | 803 | 10.35% |
| | 10th decile | 767 | 9.89% |

**Table 2.** Descriptive statistics for continuous variables. Q3: Third quartile; SD: Standard deviation

| Variable | Mean | Median | Q3 | SD | Min | Max |
|---|---|---|---|---|---|---|
| age (years) | 64.29 | 63.00 | 72.00 | 10.46 | 50.00 | 101.00 |
| Creatinine (mg/dl) | 1.09 | 1.00 | 1.20 | 0.41 | 0.40 | 10.80 |
| Kappa (mg/dl) | 1.43 | 1.27 | 1.68 | 0.89 | 0.01 | 20.50 |
| Lambda (mg/dl) | 1.70 | 1.51 | 1.92 | 1.03 | 0.04 | 26.60 |

**2.2** Data preprocessing

For the purpose of our analysis, we excluded the variable *sample.yr* which indicated the year when the measurement was recorded, and *mgus*, a binary variable (yes/no) that indicated whether the patient had been diagnosed with monoclonal gammopathy. Additionally, we removed 115 patients with MGUS, as a negative prediction in such cases was well established (12). Out the two remaining categorical variables (*sex*, *chapter*) only the binary 'sex' was used as feature in the following analysis. This variable was encoded using the LabelEncoder class in the module 'sklearn'. Since it was a binary variable, the effect was similar to using the one-hot encoder procedure. The variable *chapter* served as the target, and using label encoding for the target does not impose spurious ordinality in SVMs; this is standard practice in classification tasks (29). The missing *creatinine* values were imputed using the KNNImputer from the scikit-learn module, which estimated the missing values by analyzing the k-nearest neighbors, with known values for the feature.

The target variable was defined according to each objective, starting from the original *chapter* data. For the first objective, two distinct categories were defined: "*censored"*, representing patients with unknown survival outcomes due to right censoring, and "*died"*, including all patients who passed away during the study period. For the second objective, the target variable defined the following two categories: the "*circulatory"* class, of patients who died from circulatory-related conditions, and the "*other"* class containing all remaining patients, including both right-censored individuals and those who died from other causes.

The dataset was divided into training and testing sets using an 80/20 split. Stratification was applied based on the two classes of the target variable to maintain proportional representation across both sets. To improve model performance during training, the training set was augmented with synthetic data. Synthetic Minority Over-sampling Technique (SMOTE) (30) was employed to increase the number of minority class instances. By generating synthetic samples through interpolation between existing minority examples, SMOTE enhances the class's representation in the dataset. This technique maintains the intrinsic data structure while promoting a more balanced class distribution, which is essential for improving the accuracy and reliability of ML models. No synthetic data was generated for the testing set.

**2.3** Modeling

Three ML algorithms were used to train the classifiers for the first 2 objectives: support vector machine (SVM) (31), random forest (RF) (32), and XGBoost classifier (XGBClassifier) (33). Their performance comparison was based on the AUC value for the ROC curve of each algorithm, after parameter optimization. The set of optimized parameters were: (a) 'gamma', 'eta', and 'max_depth' for XGBClassifier; (b) 'n_estimators' and 'max_depth' for RF, and (c) 'C', 'kernel', and 'gamma' for SVM. The parameter optimization was performed using 5-fold cross validation on training data. The optimal value parameters were then used to train each model on the full training data set, and the AUC score on the test data was used for evaluation.

The third objective of the study, namely to calculate the time-dependent survival

probability for patients at risk, was achieved by using the XGBoost Survival Embeddings (XGBSE) module (34). In many cases, identifying the patients at risk represents only the first step in medical intervention, and an estimation of the probability of survival beyond a certain time is desired. This can be obtained by using survival analysis technique which accounts for the censored data.

The XGBSE package leverages XGBoost classifiers to estimate the probability of survival at a series of specified time points. To achieve this, a single survival model was first trained on the full training dataset. This trained model was then applied to the test data, generating predictions separately for three groups: patients who died from circulatory conditions, those who died from any cause, and those who were right censored. The survival predictions for these groups were then presented side by side for comparison.

We calculated the Variance Inflation Factors (VIF) for the features, and the values were all below 3.25, showing that multicollinearity was adequately assessed. However, it is worth mentioning that one important benefit of the selected ML algorithms is their robustness to data multicollinearity. This concept refers to a situation where two or more features are highly correlated, meaning one can be linearly predicted from the others with high accuracy. Multicollinearity can pose challenges in some statistical models, especially those that assume feature independence (e.g., linear regression). However, in the case of XGB, RF, and SVM, the impact of multicollinearity is considerably minimized. XGB and RF are tree-based models, and they perform feature selection during the tree-building process (25). Highly correlated features may be split on once and ignored thereafter; therefore, including redundant features does not degrade model performance. In turn, SVM contains a parameter that controls regularization and plays a critical role in balancing model complexity and error tolerance. Regularization reduces the impact of redundant or correlated features and helps with noisy and overlapping data (35). Thus, standard dimensionality reduction and feature orthogonalization (like Principal Component Analysis) are not needed for the success of the classification.

**2.4** Description of the ML algorithms

SVM is a powerful supervised ML algorithm commonly used for classification tasks. It works by finding the optimal hyperplane that best separates data points from different classes in a high-dimensional space. The goal of SVM is to maximize the margin between the closest data points of opposing classes, known as support vectors. This approach makes SVM particularly effective when there is a clear separation between classes. For cases where the data is not linearly separable, SVM can use kernel functions—such as radial basis function (RBF) or polynomial kernels—to map the input space into a higher-dimensional space where a separating hyperplane can be found.

RF is an ensemble learning method that combines the predictions of multiple decision trees to improve accuracy and reduce overfitting. It works by training each tree on a random subset of the training data and a random subset of features. Each tree makes a prediction, and the final output is determined by majority voting in classification tasks. RF is known for its

robustness, ability to handle both numerical and categorical data, and resistance to overfitting, especially in noisy datasets.

XGBClassifier is an advanced implementation of gradient boosting that has become widely popular due to its high performance, scalability, and efficiency. XGBoost builds models in a sequential manner, where each new tree aims to correct the errors made by the previous ones using a gradient descent approach to minimize a specific loss function. It includes regularization techniques (such as L1 and L2 penalties) to prevent overfitting and supports features like parallel processing and handling of missing data.

XGBSE is a survival analysis library built on top of XGBoost, designed to model the time until an event occurs (e.g., failure, death, or churn), while effectively handling censored data (i.e., where the exact timing of the event is unknown). Unlike traditional models like the Cox Proportional Hazards model, which assume linear relationships and proportional hazards, XGBSE offers a non-parametric approach. It wraps XGBoost regressors to predict survival curves without assuming any specific distribution, leveraging XGBoost's non-linear, tree-based structure for greater flexibility and performance.

Parameter optimization during the training phase by using a 5-fold cross-validation technique of the training set was performed for all algorithms.

**2.5** Evaluation metrics

Accuracy, precision, recall, and F1-score are critical performance metrics for evaluating classifiers such as XGBoost (XGBClassifier), Random Forest (RF), and Support Vector Machine (SVM), particularly in binary classification tasks. Brier score and C-statistic were used to evaluate the performance of survival analysis. In addition, the model performance was evaluated using Receiver Operating Characteristic (ROC) curves, and the Area Under the Curve (AUC) was calculated for each model.

Accuracy is one of the most used metrics to evaluate classification algorithm performance. It is defined as the ratio of correctly predicted instances to the total number of predictions. However, in the presence of imbalanced class distributions, accuracy can be misleading, as it may be disproportionately influenced by the majority class. $Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$ (eq 1), where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives. Precision quantifies the proportion of true positives (TP) among all positive predictions (TP + FP), making it valuable when false positives carry a high cost. $Precision = \frac{TP}{TP + FP}$ (eq 2) Recall, or sensitivity, measures the ability to correctly identify actual positives, calculated as TP / (TP + FN), and is essential when minimizing false negatives is a priority. $Recall = \frac{TP}{TP + FN}$ (eq 3). F1 score is the harmonic mean of precision and recall, offering a single metric that balances both, especially useful when dealing with imbalanced class distributions. $F1 = 2\frac{Precision \times Recall}{Precision + Recall}$ (eq 4)

Accuracy, precision, recall, and F1 score range from 0 to 1, with higher scores indicating better model performance. A score closer to 1 suggests that the model is making more correct predictions (accuracy), correctly identifying relevant instances (precision), capturing most of the actual positive cases (recall), and achieving a good balance between precision and recall (F1-score).

ROC curve is a graphical representation of a binary classifier's performance across different classification thresholds. It is used to evaluate how well a model distinguishes between two classes, typically *positive* and *negative*, by analyzing its true positive rate (TPR) (i.e., recall) versus its false positive rate (FPR).

Brier Score is a metric used to evaluate the accuracy of probabilistic predictions for binary outcomes. It measures the squared difference between the predicted probability of an event and the actual outcome (returning a '1' if the event occurred, and '0' if it did not). $BrierScore = \frac{1}{N}\sum_{t=1}^{N}(pred_{prob} - outcome)^2$, N= number of predictions (eq 5). This score ranges from 0 to 1, where a lower score indicates better prediction accuracy, with 0 representing a perfect prediction. The Brier Score is particularly useful for assessing how well a model's probability estimates align with the observed events, providing insight into both the model's calibration and overall performance.

C-statistic (concordance statistic), also known as the *concordance index (c-index),* is a metric used to evaluate a model's ability to correctly rank predicted outcomes in relation to observed outcomes. It is commonly applied in survival analysis and binary classification to assess the model's *discriminative power*—its ability to distinguish between individuals who experience an event and those who do not. In survival analysis, the C-statistic measures whether the predicted risk scores correctly rank individuals by their observed event times, accounting for censored data.

In binary classification, the C-statistic is equivalent to the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). A C-statistic of 0.5 indicates no better discrimination than chance, while values closer to 1.0 signify excellent predictive discrimination. Mathematically, the C-statistic (or concordance statistic) is defined as the proportion of all possible pairs of subjects where the predictions are correctly ranked concerning the observed outcomes, as follows, C- statistic = $\frac{Nc + 0.5\,Nt}{Nc + Nd + Nt}$ (eq 6). where Nc = concordant pairs, Nd = discordant pairs, Nt = tied pairs

## 3. Results and Discussion

**3.1** Model for binary classification of patient mortality, independent of condition

We compared the performance of three models (SVM, RF and XGBClassifier) so as to identify the optimal ML model that would classify patients who died during the clinical follow up and those who were right censored (Table 3). Among these models, XGBClassifier performed slightly better (accuracy= .77, precision =.81, recall = .79, F1 score= .79, AUC = .82) than the rest of the models. Consequently, the XGBClassifier was selected for further analyses.

**Table 3**. Performance evaluation of SVM, RT, and XGBClassifier classification models in distinguishing between deceased and right-censored patients during clinical follow-up.

| Model | Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| SVM | .76 | .80 | .78 | .79 | .82 |
| RF | .75 | .79 | .76 | .78 | .81 |
| XGBClassifie | .77 | .81 | .79 | .79 | .82 |

Hyperparameter tuning via grid search identified the optimal configuration as gamma= 100, eta= 0.5, and max_depth= 2, the model's best configuration. A confusion matrix was used to assess the performance of the XGBClassifier on the test dataset (Figure 2).

The model demonstrated acceptable precision (0.81), indicating its effectiveness in minimizing false positive predictions and suggesting confidence in positive classifications. However, the recall value of 79% meant that ~ one in five actual positive cases were not identified by the model, which could be a limitation in scenarios where false negatives carry significant consequences. The resulting F1-score reflected a favorable balance between precision and recall, indicating that the model achieved reliable overall performance.

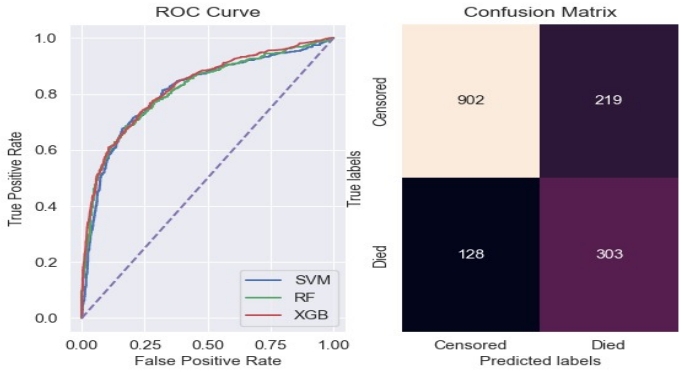


**Figure 2.** Models predicting patients survival status (*dead* vs. *right-censored* patients): ROC curve for SVM, RF, and XGB (left); Confusion matrix for XGBClassifier (right).

**3.2** Binary classification model for predicting mortality due to circulatory conditions

Three classification algorithms, Support Vector Machine (SVM), Random Forest (RF), and XGBoost Classifier (XGBClassifier), were evaluated to develop a model capable of distinguishing patients who died due to circulatory conditions and all other cases. Among these models, RF performed better (accuracy= .77, precision =.90, recall = .80, F1

score= .83, AUC = .81) than the rest of the models; therefore, it was retained for further analyses (Table 4). Hyperparameter tuning via grid search identified the optimal configuration as n_estimators= 200, and max_depth= 6, the RF model's best configuration.

**Table 4**. Performance evaluation of SVM, RT, and XGBClassifier classification models in distinguishing between the patients that died due to circulatory conditions and all other cases.

| Model | Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| SVM | .76 | .90 | .76 | .80 | .80 |
| RF | .77 | .90 | .80 | .83 | .81 |
| XGBClassifie | .76 | .90 | .77 | .81 | .80 |

A confusion matrix was used to assess the performance of RF on the test dataset (Figure 3). Overall accuracy was ~ 77%, but this was partly influenced by the higher number of censored cases. Precision was ~ 90%, indicating that some of the positive predictions (deaths) were incorrect. The F1 score, balancing precision and recall, was ~ 83%, reflecting a moderate ability to detect true deaths while managing false positives.

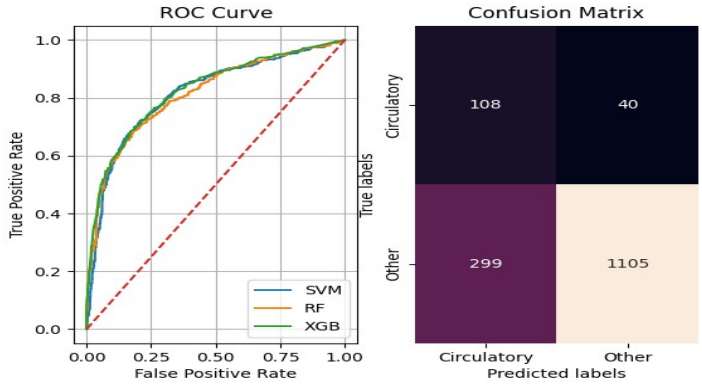


**Figure 3.** Model predicting patients who died from a circulatory condition: ROC curve for SVM, RF, and XGB (left); Confusion matrix for RF (right).

**3.3** Time-dependent survival probability model

Time-dependent survival probabilities were used to compare the survival outcomes over time across three distinct patient groups: those who died from any cause during the study period, those who died specifically from circulatory conditions, and those who survived until the end of the study.

Figure 4 illustrates how the survival probabilities changed over time for patients in each of the three groups within the test dataset.

The left panel compares patients who eventually died from any cause (red curves) to those who survived until the end of the observation period (black curves). The right panel focuses specifically on deaths attributed to circulatory conditions, again contrasting them (red curves) with patients who were right-censored (black curves). In both panels, the divergence of the curves clearly demonstrates that the model assigns lower survival probabilities to patients who died during the study compared to those who survived, indicating effective risk stratification.

The model's performance was measured in two ways. First, it was evaluated using two key metrics: C-statistic and Brier score. The C-statistic, also termed as the area under the receiver operating characteristic curve (AUC), was 0.79, indicating good discriminative ability. This showed the model was able to effectively distinguish between individuals with and without the outcome of interest, correctly ranking positive cases higher than negative ones in ~ 79% of pairs. The Brier score, which measures the accuracy of probabilistic predictions, was 0.09. This low value suggested that the predicted probabilities closely aligned with the actual outcomes, indicating that the model's probability estimates were well calibrated. Overall, these results demonstrated that the model performed well in terms of both discrimination and calibration.

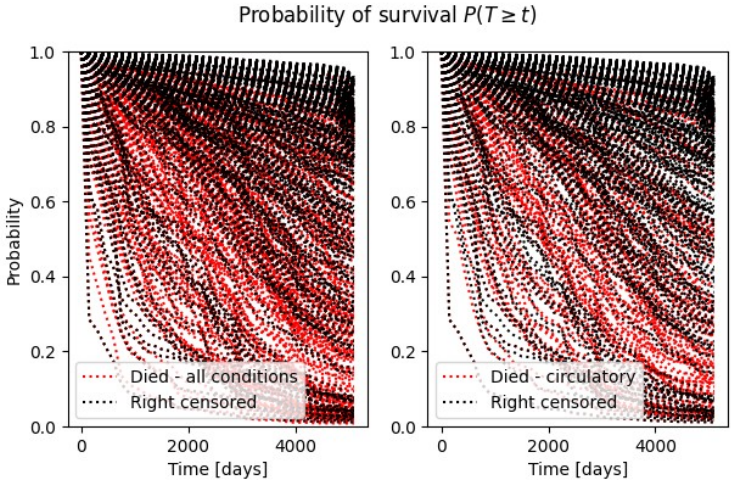


**Figure 4.** Probability of survival vs. time: (1) left: patients who died from any cause (red) compared to those who survived (black); (2) right: patients who died from circulatory conditions (red) versus those who survived (black).

Second, we computed the dynamic AUC (36) at intervals of 150 days (Figure 5) to assess the model's discriminatory ability over time. The metric quantifies how well a model can distinguish subjects who would experience an event by time t from those who would not. While standard AUC measures the discrimination for binary outcome, the dynamic AUC determines the model's discrimination power up to and including the time t. More important, it has the possibility to account for the censoring prior to time t, by including the Inverse Probability of

Censoring Weights (IPCW). Thus, this measure is not simply a calculation of standard AUCs at different times, but a more complex metric capable of accounting for patients who exited the study before experiencing the event. Essentially, the calculation was performed by considering all pairs made up of those who experienced the event by the time t and those who did not, weighted by the IPCW. As an overall measure of model discrimination, the AUC must not be interpreted as a substitute of model accuracy for predicting which patients would survive past a certain date. While AUC measures the overall ability of the model to rank a randomly chosen patient that died before a specific date higher than a randomly chosen patient that survived past that date, the prediction whether a patient will die or survive is made based on the survival probability for that specific patient.

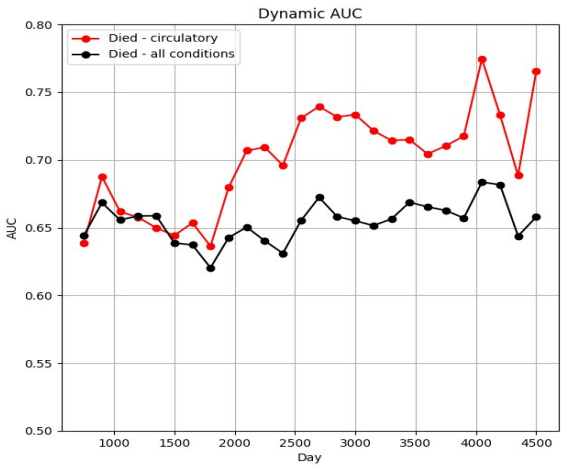


**Figure 5.** AUC for patients who died from circulatory conditions vs. patients who died from any condition.

The AUC for circulatory-related deaths remained consistently higher than that for all-cause mortality across the time period, indicating that the model was more effective at predicting deaths specifically related to circulatory issues. While both outcomes showed some variability in AUC over time, the performance for circulatory deaths exhibited a noticeable upward trend with several high points approaching or exceeding 0.75, reflecting strong discrimination ability. In contrast, the AUC for all-cause mortality remained more stable and lower, hovering around 0.65, suggesting moderate predictive power. These results highlighted the model's strength in targeting specific causes of death over more generalized outcomes.

### **3.4** Study limitations

The primary limitation of our study originated from its retrospective design. While such designs are inherently subject to various threats to internal and external validity, which may alter the interpretation and generalizability of the results (37), this challenge is common among most studies.

A second limitation was the paucity of data available for analysis. This is a common challenge in medical research, often stemming from strict patient inclusion criteria, limited availability of cases, and concerns over data privacy and sharing. Small sample sizes can reduce the statistical power of the study, increase the risk of overfitting in predictive models, and limit the ability to perform meaningful subgroup analyses. Furthermore, missing or censored data, which are frequently encountered in clinical settings, can introduce bias and reduce the generalizability of the findings (38).

A third limitation was that, although the ML models integrate comprehensive patient data to generate individualized survival predictions, such predictive approaches are still in their early stages of development (39). Consequently, no existing medical studies address the application of personalized care or pharmacological regimens based on the fine granularity of FLC levels proposed in this work.

A fourth limitation was the reliance on FLC levels as a primary biomarker. Although FLC levels provide valuable prognostic information, exclusive reliance on a single biomarker risk overlooks other potentially informative, cost-effective, or more easily measurable indicators that could be detected earlier in disease progression. Future research should aim to explore and validate complementary biomarkers that may enhance FLC's prognostic capabilities, particularly in early-stage or ambiguous cases. ML algorithms, such as the ones discussed in the present paper, can easily incorporate additional biomarkers as they become available and even determine their importance in making the prediction.

Despite these limitations, the findings of this study provide valuable preliminary insights and highlight important trends that warrant further investigation. While the current dataset may not be sufficient for highly accurate predictive modeling, it serves as a foundation for future research. Expanding the dataset through multicenter collaborations, longitudinal data collection, or integrating additional sources could significantly enhance the reliability of ML models. Future studies should prioritize data acquisition strategies to improve model performance and ensure more generalizable and clinically meaningful results.

## 4. Conclusions

ML algorithms are capable of capturing complex, nonlinear patterns in data that conventional threshold-based methods may fail to detect. As shown in Figure 4, our approach enabled personalized risk estimation by leveraging the full range of available data. Furthermore, the performance of the algorithm could be enhanced with the inclusion of additional clinical features, reflecting a key strength of ML: its capacity to integrate new data over time. For example, combining FLC measurements with omics data, bone marrow cytometry, imaging features, or longitudinal biomarker trends could generate composite risk profiles that extend beyond simple FLC cutoffs alone (40). In this way, ML can move beyond refining existing predictions to identify novel disease subtypes, forecast atypical progression trajectories, and inform trial stratification - applications that go beyond simple threshold diagnostics.

We compared the ML model with the conventional binary FLC cutoff by focusing on patients in the top decile of FLC levels, identified as highest-risk in previous work (4). Among these patients, 61% died during follow-up, meaning the standard approach would misclassify ~ 39% of survivors as "high risk." In contrast, the ML model correctly predicted 75% of deaths within this group, highlighting its ability to provide more precise, individualized risk assessments and complement traditional methods in clinical decision-making. However, these findings are based on a single comparative study and should be interpreted cautiously until validated in larger, independent cohorts.

Dynamic AUC calculation was performed at each moment of time by considering two groups of patients: those who suffered the event by that time ("cases") and patients who survived past that time ("controls"). The main difference from the standard AUC determination is that the dynamic AUC also takes into account the fact that some of the patients might have dropped out of the study until that moment (right censoring). As we did not know if the individuals who left the study would have been "cases" or "controls", the persons in the two groups above were "weighted" by the IPCW, thus ensuring that the right censored cases were accounted for. The IPCW value for each person was determined by estimating the censoring distribution from the survival times in the training data.

Difference in dynamic AUCs between the circulatory deaths and deaths due to other conditions therefore reflected each model's discriminative ability for these outcomes; it does not imply that a higher discriminative ability for deaths due to circulatory condition can be used as a marker that the death is due to this condition for each patient. Specifically, given the same patient for which death probabilities are calculated through both models, the decision of which condition has a higher chance of leading to the event is made based on the two probability values, and the physician's observations. A model with a higher dynamic AUC is expected to be more accurate in determining the individual event probabilities.

The results showed that ML algorithms effectively identified patients at higher risk of death when FLC concentrations were elevated. This indicated that changes in FLC levels were strongly linked to overall mortality. Moreover, the models could also distinguish between deaths due to circulatory causes and those from other conditions - an important capability for tailoring clinical decisions and interventions. In addition, the models generated reasonably accurate survival predictions over time. These time-dependent predictions could help estimate how long a patient might survive based on his/her clinical profile. This capability extends beyond simple risk classification, offering clinicians a powerful, informative, and practical tool for diagnosis, prognosis, and individualized care.

## 5. References

1. Di Cesare, M., Perel, P., Taylor, S., Kabudula, C., Bixby, H., Gaziano, T. A., et al. (2024). The heart of the world. *Global heart*, 19(1), 11. https://doi.org/10.5334/gh.1288

2. Basile, U., La Rosa, G., Napodano, C., Pocino, K., Cappannoli, L., Gulli, F., et al. (2019). Free light chains a novel biomarker of cardiovascular disease. A pilot study. *Eur Rev Med Pharmacol Sci*, 23(6), 2563-2569. https://www.europeanreview.org/wp/wp-content/uploads/2563-2569.pdf

3. Jackson, C. E., Haig, C., Welsh, P., Dalzell, J. R., Tsorlalis, I. K., McConnachie, A., et al. (2015). Combined free light chains are novel predictors of prognosis in heart failure. *JACC: Heart Failure*, 3(8), 618-625. https://www.jacc.org/doi/abs/10.1016/j.jchf.2015.03.014

4. Dispenzieri, A., Katzmann, J. A., Kyle, R. A., Larson, D. R., Therneau, T. M., Colby, C. L.,et al. (2012, June). Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. *Mayo Clinic Proceedings,* 87(6), 517-523. http://dx.doi.org/10.1016/j.mayocp.2012.03.009

5. Dispenzieri, A., Kyle, R., Merlini, G., Miguel, J. S., Ludwig, H., Hajek, R., et al. (2009). International Myeloma Working Group guidelines for serum-free light chain analysis in multiple myeloma and related disorders. *Leukemia*, 23(2), 215-224. https://doi.org/10.1038/leu.2008.307

6. Falk, R. H., Alexander, K. M., Liao, R., Dorbala, S. (2016). AL (light-chain) cardiac amyloidosis: a review of diagnosis and therapy. *Journal of the American College of Cardiology*, 68(12), 1323-1341. https://www.jacc.org/doi/full/10.1016/j.jacc.2016.06.053

7. Mankad, A. K., Sesay, I., Shah, K. B. (2017). Light-chain cardiac amyloidosis. *Current Problems in Cancer*, 41(2), 144-156.https://doi.org/10.1016/j.currproblcancer.2016.11.004

8. Palladini, G., Dispenzieri, A., Gertz, M. A., Kumar, S., Wechalekar, A., Hawkins, P. N., et al. (2012). New criteria for response to treatment in immunoglobulin light chain amyloidosis based on free light chain measurement and cardiac biomarkers: impact on survival outcomes. *Journal of clinical oncology*, 30(36), 4541-4549. https://doi.org/10.1200/JCO.2011.37.7614

9. Gudowska-Sawczuk, M., Mroczko, B. (2023). Free light chains κ and λ as new biomarkers of selected diseases. *International Journal of Molecular Sciences*, 24(11), 9531. https://doi.org/10.3390/ijms24119531

10. Witteles, R. M., Liedtke, M. (2021). Avoiding catastrophe: understanding free light chain testing in the evaluation of ATTR amyloidosis. *Circulation: Heart Failure*, 14(4), e008225. https://doi.org/10.1161/CIRCHEARTFAILURE.120.008225

11. Czyżewska, E., Wiśniewska, A., Waszczuk-Gajda, A., Ciepiela, O. (2021). The role of light kappa and lambda chains in heart function assessment in patients with AL amyloidosis. *Journal of Clinical Medicine*, 10(6), 1274. https://doi.org/10.3390/jcm10061274

12. Matsumori, A., Shimada, M., Jie, X., Higuchi, H., Kormelink, T. G., Redegeld, F. A. (2010). Effects of free immunoglobulin light chains on viral myocarditis. *Circulation research*, 106(9), 1533-1540. https://doi.org/10.1161/CIRCRESAHA.110.218438

13. Matsumori, A., Shimada, T., Nakatani, E., Shimada, M., Tracy, S., Chapman, N. M., et al. (2020). Immunoglobulin free light chains as an inflammatory biomarker of heart failure with myocarditis. *Clinical Immunology*, 217, 108455. https://doi.org/10.1016/j.clim.2020.108455

14. Matsumori, A. (2022). Novel biomarkers for diagnosis and management of myocarditis and heart Failure: Immunoglobulin free light chains. *21st Century Cardiol*. 2(1), 114. https://21stcenturycardiology.com/articles/novel-biomarkers-for-diagnosis-and-management-of-myocarditis-and-heart-failure-immunoglobulin-free-light-chains.pdf

15. Deng, X., Crowson, C. S., Rajkumar, S. V., Dispenzieri, A., Larson, D. R., Therneau, T. M., et al. (2015). Elevation of serum immunoglobulin free light chains during the preclinical period of rheumatoid arthritis. *The Journal of rheumatology*, 42(2), 181-187. https://doi.org/10.3899/jrheum.140543

16. Ren, W., Zhang, Z., Wang, Y., Wang, J., Li, L., Shi, L., et al. (2025). Coronary health index based on immunoglobulin light chains to assess coronary heart disease risk with machine learning: a diagnostic trial. *Journal of Translational Medicine*, 23(1), 22. https://doi.org/10.1186/s12967-024-06043-4

17. Libby, P., Ridker, P. M., Maseri, A. (2002). Inflammation and atherosclerosis. *Circulation*, 105(9), 1135-1143. https://doi.org/10.1161/hc0902.104353

18. Januzzi, J. L., Mann, D. L. (2015). Light chains and the failing heart: important mechanistic link or fifty shades of gray?. *JACC: Heart Failure*, 3(8), 626-628. https://www.jacc.org/doi/full/10.1016/j.jchf.2015.03.012

19. Perrone, M. A., Pieri, M., Marchei, M., Sergi, D., Bernardini, S., Romeo, F. (2019). Serum free light chains in patients with ST elevation myocardial infarction (STEMI): A possible correlation with left ventricle dysfunction. *International Journal of Cardiology*, 292, 32-34. https://doi.org/10.1016/j.ijcard.2019.06.016

20. Shantsila, E., Wrigley, B., Lip, G. Y. H. (2014). Free light chains in patients with acute heart failure secondary to atherosclerotic coronary artery disease. *The American Journal of Cardiology*, 114(8), 1243-1248. https://doi.org/10.1016/j.amjcard.2014.07.049

21. Tahir, U. A., Doros, G., Kim, J. S., Connors, L. H., Seldin, D. C., Sam, F. (2019). Predictors of mortality in light chain cardiac amyloidosis with heart failure. *Scientific reports*, 9(1), 8552.https://doi.org/10.1038/s41598-019-44912-x

22. Donnelly, J. P., Hanna, M. (2017). Cardiac amyloidosis: An update on diagnosis and treatment. *Cleve Clin J Med*, 84(12 Suppl 3), 12-26. https://doi.org/10.3949/ccjm.84.s3.02

23. Shomanova, Z., Ohnewein, B., Schernthaner, C., Höfer, K., Pogoda, C. A., Frommeyer, G., et al. (2020). Classic and novel biomarkers as potential predictors of ventricular arrhythmias and sudden cardiac death. *Journal of Clinical Medicine*, 9(2), 578. https://doi.org/10.3390/jcm9020578

24. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1), 195. https://doi.org/10.1186/s12916-019-1426-2

25. Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning: Data mining, inference and prediction. 2rd ed. Springer.

26. Maeng, C.V., Rögnvaldsson, S., Long, T. E., Brieghel, C., Hermensen, E., Niemann, C. U., *et al.* (2025).Revised free light chain reference intervals enhance risk stratification in monoclonal gammopathy of undetermined significance and reduce overdiagnosis. *Blood Cancer J.* 15, 80. https://doi.org/10.1038/s41408-025-01289-7

27. Jackson, C. E., Haig, C., Welsh, P., Dalzell, J. R., Tsorlalis, I. K., McConnachie, A., et al. (2015). Combined free light chains are novel predictors of prognosis in heart failure. *JACC: Heart Failure*, 3(8), 618-625. https://www.jacc.org/doi/10.1016/j.jchf.2015.03.014

28. Ren, W., Zhang, Z., Wang, Y., Wang, J., Li, L., Shi, L., et al. (2025). Coronary health index based on immunoglobulin light chains to assess coronary heart disease risk with machine learning: a diagnostic trial. *Journal of Translational Medicine*, 23(1), 22. https://doi.org/10.1186/s12967-024-06043-4

29. Bisong, E. (2019). Introduction to Scikit-learn. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress. https://doi.org/10.1007/978-1-4842-4470-8_18

30. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357. https://doi.org/10.1613/jair.953

31. Boser, B. E., Guyon, I. M., Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). https://dl.acm.org/doi/pdf/10.1145/130385.130401

32. Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32. https://link.springer.com/content/pdf/10.1023/a:1010933404324.pdf

33. Chen, T., Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). http://dx.doi.org/10.1145/2939672.2939785

34. Vieira, D., Gimenez, G., Marmorela, G., Estima, V. (2021). XGBoost survival embeddings: Improving statistical properties of XGBoost survival analysis implementation. *Loft Python*. https://github.com/loft-br/xgboost-survival-embeddings

35. Xu, H., Caramanis, C., Mannor, S. (2009). Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 1485-1510. https://www.jmlr.org/papers/volume10/xu09b/xu09b.pdf

36. Hung, H., Chiang. C. T. (2010). "Estimation methods for time-dependent AUC models with survival data," *Canadian Journal of Statistics,* 38(1), 8–26. https://doi.org/10.1002/cjs.10046

37. Trochim, W. M. (2005). *Research methods: The concise knowledge base*. Mason, Ohio: Thomson. https://cir.nii.ac.jp/crid/1971712334825297280

38. Dash, S., Shakyawar, S. K., Sharma, M., Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of big data*, 6(1), 1-25. https://doi.org/10.1186/s40537-019-0217-0

39. Durso-Finley, J., Barile, B., Falet, J. P., Arnold, D. L., Pawlowski, N., Arbel, T. (2024, October). Probabilistic temporal prediction of continuous disease trajectories and treatment effects using neural SDEs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 400-410). Cham: Springer Nature Switzerland. https://arxiv.org/pdf/2406.12807

40. Mateos, M. V., González-Calle, V. (2017). Smoldering multiple myeloma: Who and when to treat. *Clinical Lymphoma Myeloma and Leukemia*, 17(11), 716-722. https://doi.org/10.1016/j.clml.2017.06.022