Scientific Journal of King Faisal University:
Basic and Applied Sciences

Artificial Intelligence and Predictive Medicine: Predicting Thrombocytopenia in Patients with Atrial Fibrillation

Amadou Diabagate [1], Katienefowa Sekou Koulibaly [2], Awa FOFANA [2] and Doffou Jérôme Diako [3]

[1] Faculty of Mathematics and Computer Science, University Felix Houphouët-Boigny, Cocody, Abidjan, Côte d'Ivoire
[2] Faculty of Medicine, University Felix Houphouët-Boigny, Abidjan, Côte d'Ivoire
[3] Ecole Supérieure Africaine des TIC (ESATIC), Abidjan, Côte d'Ivoire

LINK	RECEIVED	ACCEPTED	PUBLISHED ONLINE	ASSIGNED TO AN ISSUE
https://doi.org/10.37575/b/med/250029	12/08/2025	30/11/2025	30/11/2025	01/12/2025

NO. OF WORDS	NO. OF PAGES	YEAR	VOLUME	ISSUE
8439	10	2025	26	2

ABSTRACT

This study aimed to develop and evaluate machine learning models capable of predicting thrombocytopenia in patients with atrial fibrillation, with a focus on African clinical settings where this complication is underexplored despite its clinical importance. Real-world data were obtained from the AFRICA registry, a large multicenter database encompassing diverse patient profiles. Six algorithms were implemented and compared, including Decision Tree, Random Forest, XGBoost, Support Vector Machine, K-Nearest Neighbors, and Multi-Layer Perceptron. Stratified cross-validation was used to ensure robust evaluation based on accuracy, F1-score, AUC-ROC, Log Loss, and Matthews Correlation Coefficient. Model interpretability was enhanced using the SHAP method to identify the most influential predictors. Tree-based models performed best. On cross-validation, XGBoost reached 97.4%, F1 0.92, and AUC-ROC 0.98, and its performance was confirmed on an independent 20% holdout set (accuracy 93.75%, F1 0.857, AUC-ROC 0.9803). SHAP analysis highlighted platelet count, hemoglobin, creatinine, and glycemia as the strongest predictors, alongside clinical factors such as amiodarone therapy, intensive care admission, and depressive symptoms. High-performance, interpretable machine learning models can accurately forecast thrombocytopenia in African patients with atrial fibrillation. These findings provide a solid basis for the development of clinical decision support systems aimed at improving patient management and treatment outcomes.

KEYWORDS

AFRICA registry, clinical decision, hematology, model interpretability, multicenter cohort, supervised learning

CITATION

Diabagate, A., Koulibaly, K.S., Fofana, A. and Diako, D.J. (2025). Artificial intelligence and predictive medicine: Predicting thrombocytopenia in patients with atrial fibrillation. *Scientific Journal of King Faisal University: Basic and Applied Sciences*, **26**(2), 69–78. DOI: 10.37575/b/med/250029

1. Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia globally, affecting over 37 million individuals according to recent estimates (Hindricks *et al.*, 2021). It is a major contributor to adverse cardiovascular outcomes, including stroke, heart failure, cognitive decline, and premature death (Chugh *et al.*, 2014). While AF has been extensively studied in high-income countries, data from sub-Saharan Africa remain limited. However, rapid demographic and epidemiological transitions are increasing the burden of AF across the continent (Ding and Tang, 2025; Dzudie *et al.*, 2020). Thrombocytopenia, an abnormally low platelet count, poses a serious clinical concern for AF patients. It compromises the ability to maintain optimal anticoagulation therapy, which is crucial for preventing thromboembolic complications such as stroke (Iijima *et al.*, 2023; Yeh *et al.*, 2022). The causes of thrombocytopenia in this population are multifactorial and include comorbid conditions (e.g., hypertension, renal impairment), pharmacologic agents (notably anticoagulants), and systemic inflammatory or infectious states (Iijima *et al.*, 2023; Yeh *et al.*, 2022). Despite its clinical importance, early prediction of thrombocytopenia in AF patients has been largely neglected in the literature, particularly in African settings where access to advanced biomarkers is often limited. Machine learning (ML) methods offer a compelling opportunity to address this gap. Unlike traditional statistical approaches, ML algorithms are well suited for handling heterogeneous clinical datasets, capturing complex nonlinear relationships, and managing imbalanced class distributions, common features of medical data (Haibo and Garcia, 2009; Rajkomar *et al.*, 2019). Furthermore, the integration of explainability tools, such as SHAP (SHapley Additive exPlanations), strengthens model transparency and clinical interpretability (Lundberg and Lee, 2017). This study applies an integrated approach combining machine learning, model explainability, and African clinical data to develop a robust predictive model for thrombocytopenia in AF patients. We leveraged data from the AFRICA registry, a structured pan-African initiative for documenting AF cases, implemented several advanced algorithms (XGBoost, Decision Tree, Random Forest, Support Vector Machine, Neural Network and K-Nearest Neighbors), used SMOTE for class balancing, applied stratified cross-validation, and conducted SHAP analysis to interpret model decisions. This research introduces a novel approach that brings together predictive modeling and interpretable Artificial Intelligence (AI), specifically applied to African AF cohorts for the first time. The proposed methodology is particularly relevant for resource-limited healthcare systems, where accurate yet interpretable decision support is essential to improving clinical outcomes. The paper is organized into eight sections. Section II explores the clinical intersection between AF and thrombocytopenia and introduces the AFRICA registry. Section III presents a review of the relevant literature. Section IV outlines the methodology. Section V details the results, followed by a critical discussion in Section VI. Finally, Section VII concludes the study and outlines future research directions.

2. Literature Review

The clinical management of atrial fibrillation (AF) relies heavily on anticoagulation to prevent thromboembolic events, but this strategy is complicated when patients develop hematologic disorders such as thrombocytopenia. This condition significantly increases the risk of thromboembolic complications, which justifies the widespread use of oral anticoagulants, either Direct Oral Anticoagulants (DOACs) or Vitamin K Antagonists (VKAs), as a key element of stroke prevention strategies (January *et al.*, 2019; Moulis *et al.*, 2020). However, these therapies also elevate the risk of bleeding, particularly in patients with underlying hematological abnormalities. Thrombocytopenia, defined as a reduced number of circulating platelets, is frequently observed in patients receiving anticoagulants. Its origins are multifactorial and may include drug-induced toxicity, immune-mediated responses, bone

Corresponding Author: Amadou Diabagate 002250767954216, amadou1.diabagate@ufhb.edu.ci

marrow suppression, or comorbid conditions such as cancer, autoimmune diseases, or chronic infections (Stasi, 2012). In patients with AF, the emergence of thrombocytopenia often necessitates reassessment of the anticoagulation regimen, adjusting the dosage or even temporarily discontinuing treatment, thereby exposing the patient to an increased risk of stroke (Lip *et al.*, 2022). Several studies have established a link between thrombocytopenia and worse clinical outcomes in AF, including a higher risk of major bleeding, longer hospital stays, and more complex therapeutic decisions. Yet, despite these risks, thrombocytopenia remains poorly investigated in AF cohorts. Common bleeding risk scores, such as HAS-BLED and ORBIT, which are clinical tools designed to estimate the likelihood of bleeding in patients undergoing anticoagulation therapy, do not consistently include platelet count as a predictive variable (Pisters *et al.*, 2010). Early identification of AF patients at high risk of developing thrombocytopenia could support more personalized anticoagulant management, allowing for closer monitoring, dose adjustments, or early exploration of underlying causes. In this regard, advanced modeling techniques such as machine learning (ML) offer promising potential. ML algorithms can simultaneously handle a large number of clinical and biological variables, detect complex interactions, and improve predictive performance beyond traditional statistical methods (Rahul *et al.*, 2025). The goal of this study is to build and evaluate a classification model based on multiple ML algorithms capable of predicting thrombocytopenia in patients with AF using structured clinical data. To enable such data-driven approaches in African populations, robust clinical datasets are essential. The AFRICA Registry (Atrial Fibrillation Registry in Countries of Africa) was established by a collective of African cardiologists to address the scarcity of reliable epidemiological data on atrial fibrillation in sub-Saharan Africa, where the growing burden of cardiovascular disease remains poorly documented. Demographic shifts, urbanization, and changing health profiles have contributed to a rise in AF cases across the continent, yet most existing studies remain limited in scope and generalizability (Mendis *et al.*, 2011). The AFRICA Registry is a prospective, multicenter study designed to systematically collect clinical, biological, and therapeutic data on patients with AF. Its objectives include characterizing patient profiles, documenting treatment practices, monitoring adverse events such as stroke and bleeding, and evaluating the applicability of global risk stratification tools in African healthcare contexts (Diop *et al.*, 2022). Eligible patients are adults diagnosed with AF, recruited from hospitals across both urban and rural settings. Data are collected through harmonized electronic forms and stored in a centralized, secure database aligned with international clinical research standards. Preliminary results from countries such as Côte d'Ivoire, Senegal, and Cameroon show one-year mortality rates reaching 48%, largely due to ischemic complications and heart failure (Ettarh, 2016). The data also reveal inconsistencies in care delivery, including underuse of anticoagulants and delayed diagnoses, emphasizing the need for region-specific treatment guidelines (Stambler and Ngunga, 2015). From a methodological perspective, the registry offers a unique opportunity to apply artificial intelligence and machine learning techniques to real-world African data, supporting the development of predictive models for personalized patient care (Liu *et al.*, 2025). Looking ahead, the registry aims to expand geographically and integrate additional modules addressing social determinants of health, treatment adherence, and access to cardiovascular technologies. These initiatives are intended to improve the quality, equity, and digital transformation of cardiovascular care across the continent. Atrial fibrillation (AF) is consistently associated with stroke, heart failure, and excess mortality, making early and accurate risk stratification a major clinical priority (Chugh *et al.*, 2014; January *et al.*, 2019; Stambler and Ngunga, 2015). Rhythm or rate control is typically combined with oral anticoagulation to prevent thromboembolic events, but this strategy inevitably increases the risk of bleeding, especially when hematological abnormalities such as thrombocytopenia coexist (Hindricks *et al.*, 2021; Verheugt and Granger, 2015). Traditional prediction tools including HAS-BLED remain useful in clinical practice, yet they often overlook biologically relevant markers such as platelet indices, which may reduce their performance in specific patient subsets (Iijima *et al.*, 2023; Pisters *et al.*, 2010; Rahul *et al.*, 2025). Machine learning (ML) has progressively demonstrated superior predictive capabilities over conventional scores for AF-related outcomes by leveraging richer clinical and biological inputs (Lu *et al.*, 2024). Nevertheless, most existing AF-focused models still concentrate on global bleeding risk rather than platelet decline itself (Rahul *et al.*, 2025). Conversely, ML-based thrombocytopenia prediction has been investigated primarily in non-AF contexts such as postoperative care, sepsis, or intensive care settings, where approaches like Random Forest, XGBoost or Support Vector Machines have shown promising performance (Cheng *et al.*, 2021; Jiang *et al.*, 2022). In hematology, ML has also contributed to diagnostic support and treatment planning in immune thrombocytopenia, illustrating its value for platelet-related disorders (Ghanima and Cooper, 2024; Haroon *et al.*, 2024). Reviews have also clarified the mechanisms underlying thrombocytopenia, reinforcing the biological plausibility of platelet-centric risk modeling in AF populations (Stasi, 2012). To situate the current study within this evidence base, Table 1 synthesizes representative work on AF risk prediction and thrombocytopenia modeling. This comparison underscores the absence of ML approaches designed to predict thrombocytopenia specifically within AF cohorts, including in African clinical environments where early detection could help tailor safer anticoagulant strategies.

Table 1. Comparative overview of related studies on AF risk prediction and thrombocytopenia modeling Comparative summary of AF risk-prediction and thrombocytopenia-modeling studies, highlighting populations, endpoints, methods, and the lack of AF-specific thrombocytopenia prediction (especially in Africa)

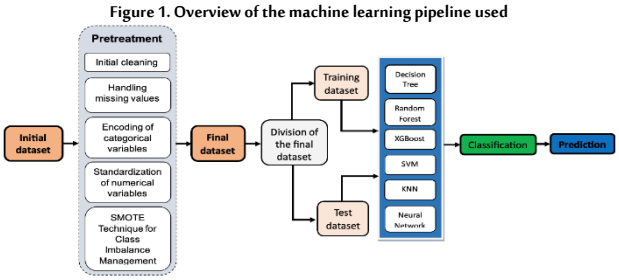
Study (Year)	Population / Context	Task / Outcome	Method (s)	Key Findings	Gap vs. This Work
Lu *et al.*, 2024	AF registry, multicenter	Predict thromboembolic and bleeding risks in AF	ML ensembles	ML improves clinical risk stratification vs. traditional scores	Does not target thrombocytopenia as outcome
Rahul *et al.*, 2025	AF patients on DOAC	Predict bleeding risk	ML models	Strong bleeding prediction with ML	Different endpoint from platelet decline
Iyengar *et al.*, 2023	AF cohorts	Association between thrombocytopenia and adverse events	Statistical analyses	Thrombocytopenia associated with bleeding and vascular events	No ML prediction task
Iijima *et al.*, 2023	AF patients with stable CAD (rivaroxaban)	Association between thrombocytopenia and bleeding in AF	Statistical analyses	Thrombocytopenia independently increases major bleeding risk	No ML prediction task
Cheng *et al.*, 2021	ICU, postoperative	Predict hospital-acquired thrombocytopenia	RF, SVM	ML feasible and informative in ICU	Not AF-specific
Jiang *et al.*, 2022	ICU, sepsis	Predict sepsis-associated thrombocytopenia	ML model	Good discrimination for SATP	Different mechanisms from AF context
Haroon *et al.*, 2024	Hematology (ITP)	Assist in ITP diagnosis	ML feasibility	ML promising for platelet disorders	Not AF; different endpoint
Ghanima and Cooper, 2024	Hematology	Perspectives on ML in ITP	Expert commentary	Underlines ML's potential for platelets	No empirical modeling
Stasi, 2012	Review	Mechanisms and clinical update	Review	Biological rationale for platelet-centric risks	Contextual, not predictive
Pisters *et al.*, 2010; Verheugt and Granger, 2015	AF management	Stroke/bleeding risk scoring	Clinical scores	Standard of care references	Do not predict thrombocytopenia
Hindricks *et al.*, 2021; January *et al.*, 2019	Guidelines	AF therapeutic strategy	Guidelines	Management frameworks for AF	No ML, no platelet-decline focus
This study	AF patients in Côte d'Ivoire (n = 239)	Predict AF-associated thrombocytopenia	XGBoost with SHAP	Strong predictive performance with interpretability	First AF-specific thrombocytopenia model in an African setting

Abbreviations: AF, atrial fibrillation; OAC, oral anticoagulant; DOAC, direct oral anticoagulant; CAD, Coronary Artery Disease; ICU, intensive care unit; ITP, immune thrombocytopenia; RF, Random Forest; SVM, Support Vector Machine; ; ML, Machine Learning.

3. Methodology

This section introduces the methodological approach adopted to predict thrombocytopenia in patients with atrial fibrillation. The proposed workflow follows a supervised machine learning pipeline that includes data preprocessing, class balancing, model training, and predictive evaluation. Figure 1 below provides an overview of the system's architecture, illustrating the overall workflow.

Figure 1. Overview of the machine learning pipeline used



To operationalize this workflow, the pseudocode below details each technical stage.

Pseudocode — Machine Learning Workflow for Thrombocytopenia Prediction

Input:

Structured clinical dataset of atrial fibrillation (AF) patients

Output:

Thrombocytopenia classification and interpretability insights

Step 1: Import essential libraries and packages

Step 2: Load the dataset and remove

- Irrelevant columns
- Columns with too many (>100) missing values

Step 3: Handle missing data

Step 4: Encode categorical variables and normalize numerical features

Step 5: Address class imbalance using SMOTE

- Apply oversampling only on the training set
- Set target minority-to-majority ratio (e.g., 0.75)f

Step 6: Split dataset into training (80%) and testing (20%)

Step 7: Define candidate machine learning models

- Decision Tree, Random Forest, XGBoost, SVM, KNN, Neural Network

Step 8: Perform hyperparameter tuning using RandomizedSearchCV

- Identify optimal configurations for each model

Step 9: Train and validate models using Stratified K-Fold Cross-Validation

Step 10: Evaluate performance on key metrics

- Accuracy, F1-score, Precision, Recall, AUC-ROC, MCC, Log-loss

Step 11: Interpret best-performing model using SHAP

- Identify most influential features
- Visualize variable impacts and directionality

3.1. Clinical Dataset Overview:

The dataset employed in this study originates from the AFRICA registry (Atrial Fibrillation Registry in Countries of Africa), a large-scale continental initiative designed to standardize the collection of clinical data on atrial fibrillation (AF) across Africa. For the purposes of this research, data were drawn specifically from the Ivorian cohort of the registry. This extraction covers two time periods: from January 1, 2016 to January 31, 2018, and from January 2021 to December 2023. This dataset was compiled through a multicenter study conducted at several major healthcare institutions in Côte d'Ivoire, including the Institut de Cardiologie d'Abidjan (ICA), Centre Médical Cardio-Respiratoire des Jardins de Cocody (CMCARE), Home Medical Service du Plateau (HMS), Clinique Saintes Myriades de Marcory, Polyclinique Internationale Sainte Anne-Marie (PISAM), Hôpital Militaire d'Abidjan (HMA), Centre Hospitalier Universitaire (CHU) de Yopougon and CHU de Bouaké. Eligible participants were adult patients (aged 18 and above) residing in Côte d'Ivoire, with a confirmed diagnosis of AF documented by either standard electrocardiogram (ECG) or Holter ECG monitoring. Only individuals with accessible clinical follow-up for a minimum of 12 months and who had provided informed consent were included. Exclusion criteria comprised pregnant women, patients not formally enrolled in the registry, and cases of situational or transient AF.

3.1.1. Description of Collected Data and Patient Profiles

The dataset offers a comprehensive and structured view of patients with atrial fibrillation, covering baseline characteristics and clinical outcomes over at least one year. Data were collected prospectively during consultations and digitally recorded using standardized forms to ensure consistency and longitudinal accuracy.

Recorded variables span demographics (age, sex, location), AF type and symptoms (EHRA score), stroke and bleeding risks ($CHADS_2$, CHA_2DS_2-VASc, HAS-BLED), and comorbidities (e.g., hypertension, diabetes, thrombocytopenia). Treatments include antiarrhythmic drugs, anticoagulants, and cardioversion methods. Outcomes such as complications, hospitalizations, and death were tracked, alongside ECG, echocardiography, and lab results. Table 2 summarizes all variables and their codings.

Table 2. Overview of key variables in the dataset the main demographic, clinical, therapeutic, and laboratory variables collected from 239 atrial fibrillation patients. It highlights the dataset's diversity and the relevance of comorbid conditions for model development

Data Type	Name	Coding	Description
Demographics	Age	Numeric (years)	Age of the patient at the time of atrial fibrillation diagnosis.
	Sex	1 = Male, 2 = Female	Biological sex of the patient.
	Location	Text (City or Region)	Geographic area where the patient received care.
Clinical	Type of Atrial Fibrillation	1 = Paroxysmal, 2 = Persistent, 3= Long-standing persistent, 4 = Permanent	Classification based on the duration and stability of the arrhythmia.
	Symptoms	1 = Palpitations, 2 = Dyspnea, 3 = Fatigue, 4 = Syncope	Main symptoms reported by patients at presentation.
	EHRA Score	1 = None, 2 = Mild, 3 = Severe, 4 = Disabling	Functional classification of symptom burden.
	$CHADS_2$ Score	Numeric (0–6)	Score estimating stroke risk based on comorbidities.
	CHA_2DS_2-VASc Score	Numeric (0–9)	Extended stroke risk score including additional clinical factors.
	Comorbidities	1 = Hypertension, 2 = Diabetes, 3 = Heart Failure, 4 = Thrombocytopenia, etc.	Coexisting medical conditions that impact clinical risk.
	HAS-BLED Score	Numeric (0–9)	Score estimating bleeding risk under anticoagulation therapy.
Therapeutics	Type of Treatment	1 = Flecainide, 2 = Amiodarone, 3 = Anticoagulants, etc.	Pharmacological strategies used for AF management.
	Cardioversion Method	1 = Pharmacological, 2 = Electrical	Method employed to restore sinus rhythm.
Outcomes	Complications	1 = Stroke, 2 = Thromboembolism, 3 = Bleeding, etc.	Adverse events occurring during the follow-up period.

Data Type	Name	Coding	Description
	Death	0 = No, 1 = Yes	Indicates whether the patient died within one year of follow-up.
	Cause of Death	Text (Narrative)	Specific clinical reason for death, when applicable.
	Hospitalizations	Numeric (Count)	Total number of hospital admissions recorded during the study.
Paraclinical Tests	ECG Results	Text or Code	Electrocardiogram findings documenting AF or related abnormalities.
	Echocardiography	Numeric and/or Text	Echographic parameters including structural heart measures and thrombus status.
Biological	Laboratory Values	Numeric	Blood test results used to support diagnosis and treatment decisions.

The descriptive analysis reveals marked clinical and biological diversity within the cohort, justifying the use of advanced modeling. Table 3 summarizes key variables. Platelet counts range widely (34,000 to 408,000/mm³; mean: 332,100), pointing to varying hematologic states, including possible thrombocytopenia. Hemoglobin also spans a broad spectrum (6.2 to 46.2 g/dL), though the median remains normal.

Serum creatinine values show significant dispersion (up to 111 mg/L), suggesting renal dysfunction in some cases. In contrast, blood glucose appears more stable (mean: 0.97 g/L). Echocardiographic metrics like LVEDD (54.2 mm) and LVESD (38.9 mm) show moderate variability, while resting heart rate (mean: 83.9 bpm, SD: 23.5) reflect clinical heterogeneity in disease expression or treatment response.

Patient ages range from 12 to 91, with a median of 65, capturing both early and late AF onset. Weight (67.8 kg) and height (167.9 cm) remain consistent with adult West African norms.

Table 3. Summary of descriptive statistics for key clinical variables central tendency and dispersion (mean, variance, SD, median, mode, range, min to max) across hematologic, metabolic, echocardiographic, and vital sign measures. Highlights wide variability in platelet and creatinine values supporting flexible modeling approaches

Variable	Mean	Variance	Standard Deviation	Median	Mode	Range	Min	Max
Platelets	$33.21.10^3$	7.500E+9	$86.60.10^3$	$253.5.10^3$	253.50	$407.97.10^3$	34.00	408.00
Hemoglobin	13.05	9.15	3.03	13.00	13.00	40.00	6.20	46.20
Serum Creatinine	13.37	96.95	9.85	12.00	12.00	110.20	0.80	111.00
Blood Glucose	0.97	0.08	0.28	0.93	0.93	3.60	0.31	3.91
Left Ventricular Mass	48.28	50.47	7.10	48.00	48.00	68.00	6.00	74.00
Left Ventricular End-Diastolic Diameter	54.20	82.74	9.10	54.00	54.00	86.00	5.00	91.00
Left Ventricular End-Systolic Diameter	38.92	65.65	8.10	38.00	38.00	60.00	10.00	70.00
Heart Rate	83.87	552.84	23.51	80.50	80.50	170.00	6.00	176.00
Previous Rhythm History	$7.55.10^2$	$1.76.10^6$	$1.33.10^3$	$3.65.10^2$	$3.65.10^2$	$9.99.10^3$	0.00	$9.99.10^3$
Age	61.74	248.89	15.78	65.00	67.00	79.00	12.00	91.00
Weight	67.84	180.72	13.44	68.00	68.00	120.00	10.00	130.00
Height	167.89	34.36	5.86	168.00	168.00	45.00	150.00	195.00

These descriptive findings illustrate the complexity and variability of patient profiles, justifying the use of machine learning approaches capable of managing non-linear interactions, heterogeneous distributions, and multiscale clinical patterns.

3.1.2. Correlation Matrix Analysis of Key Predictive Features

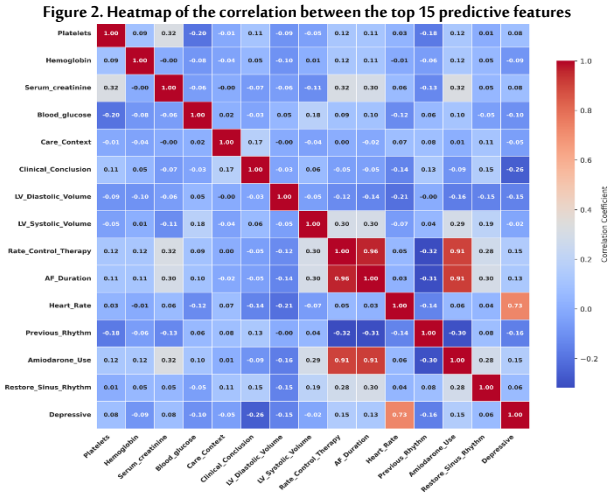
The correlation matrix of the 15 most important variables identified by the Random Forest model reveals several clinically meaningful associations that support the robustness of the feature selection process.

A moderate negative correlation was found between platelet count, hemoglobin levels, and the likelihood of sinus rhythm restoration. This suggests that patients with lower values, which often reflect frailty or comorbidities, may be less likely to undergo rhythm control strategies. Serum creatinine was mildly associated with ventricular volumes, supporting the link between renal dysfunction and cardiac remodeling.

The care setting showed strong associations with both diastolic and systolic measurements, implying that patients managed in more structured environments tend to present with more advanced cardiac profiles. The strong correlation between rate control therapy and heart rate confirms the coherence of treatment choices. Additionally, although weaker, correlations between depressive symptoms and therapies such as amiodarone suggest that psychological status may influence clinical decisions.

These observations highlight the clinical relevance of the selected features and support the model's ability to integrate physiological and contextual risk factors.

Figure 2. Heatmap of the correlation between the top 15 predictive features



3.2. Data Preprocessing and Ethical Compliance:

Before model training, we set up a rigorous preprocessing pipeline to ensure data quality and consistency. Variables with excessive missingness or limited predictive value were reviewed with clinicians and removed when appropriate. For the retained features, categorical variables were numerically encoded, and continuous variables were normalized to support convergence.

Of the 130 variables, 115 contained missing values (0.4% to 62.8%), and 91 (70.0%) had less than 25% missingness. We predefined a per-variable threshold of at least 40% missingness (about 100 of 239 observations) to trigger clinical review. Variables above this threshold were excluded only when judged non-essential; essential variables were kept. Missing values were then imputed within the training folds to avoid data leakage, using the mean for numerical variables and the mode for categorical variables.

The cohort included 53 patients with thrombocytopenia (22.2%), indicating class imbalance. We therefore applied SMOTE only to the training set after an 80/20 stratified split. This procedure generated synthetic minority samples by interpolating among existing cases, helping preserve data structure and limit overfitting. We used an oversampling ratio of 0.75, increasing thrombocytopenic cases from 42 to 105 in the training set while maintaining 140 majority cases. A brief sensitivity check at 0.5 and 1.0 produced the same model ranking, suggesting stable results across sampling choices. The test set was left untouched to reflect real-world distributions.

This approach improved class balance and enhanced sensitivity to high-risk cases, which is critical in clinical prediction where false negatives carry serious consequences. The final analysis dataset was

entirely numerical and free of missing values at the time of model training and evaluation.

All procedures complied with the Declaration of Helsinki and institutional requirements. The protocol was authorized by the Medical and Scientific Directorate of the Institut de Cardiologie d'Abidjan (Authorization No. AKJB/GK N°037-2018/MSHP/ICA/DG/DMS). All data were anonymized prior to analysis to protect patient confidentiality.

3.3. Machine Learning Methods:

To develop a robust and clinically meaningful predictive model, this study examined six supervised machine learning algorithms that are widely applied in medical data science. These methods represent distinct modeling approaches including tree-based, ensemble, kernel-based, distance-based, and neural network techniques. Their selection aimed to ensure both methodological diversity and relevance to structured clinical datasets commonly encountered in healthcare (Mendis *et al.*, 2011; Rajkomar *et al.*, 2019).

Random forest, introduced by Breiman (2001), enhances the performance of decision trees by creating an ensemble of multiple trees trained on random data subsets. The final prediction is based on majority voting across trees, which improves generalizability and reduces variance. This method also enables reliable estimation of feature importance, making it particularly valuable in clinical research where model explainability is essential.

Multilayer Perceptron (MLP) represents a classical feedforward neural network architecture composed of multiple interconnected layers of neurons (Bishop, 1995). It is capable of modeling complex nonlinear relationships and is trained using backpropagation algorithms. Despite its potential, the lack of interpretability and its dependency on large, balanced datasets pose challenges for its use in clinical practice, where transparency is a key requirement (Miotto *et al.*, 2018).

Decision Trees are among the most intuitive and transparent classification techniques. They operate by recursively partitioning the data based on impurity measures such as entropy or the Gini index, resulting in a tree-like structure that can be interpreted easily by clinicians (Breiman *et al.*, 2017). Although decision trees may overfit small datasets, their limitations can be addressed by restricting tree depth and applying post-pruning procedures.

Support Vector Machines (SVMs) aim to identify the optimal hyperplane that maximizes the separation between classes (Cortes and Vapnik, 1995). While SVMs perform well in scenarios where data are linearly separable, they often show limited adaptability in real-world clinical datasets that contain noise, class imbalance, or overlapping distributions (Cheng *et al.*, 2021). These constraints may hinder their effectiveness in complex patient populations.

XGBoost (eXtreme Gradient Boosting) is a high-performance gradient boosting algorithm developed by Chen and Guestrin (2016). It introduces advanced regularization, parallel processing, and automated pruning, resulting in superior predictive accuracy and robustness. Its effectiveness has been demonstrated repeatedly in structured medical datasets, especially under conditions of class imbalance and noise.

K-Nearest Neighbors (KNN) is a non-parametric technique that assigns class labels based on the majority class among the closest data points in the feature space. Although easy to implement, KNN is sensitive to the curse of dimensionality and tends to perform poorly in high-noise environments (James *et al.*, 2013). In this study, it produced a high rate of false positives, limiting its clinical applicability.

In summary, the comparative evaluation of these models allows for a balanced assessment of their predictive capabilities, interpretability, and operational feasibility in the context of thrombocytopenia prediction among patients with atrial fibrillation.

3.4. Hyperparameter Optimization and Model Selection:

This study evaluated several machine learning algorithms for predicting thrombocytopenia using 239 patient records and 130 clinical variables. Hyperparameter tuning was performed with RandomizedSearchCV to efficiently explore the parameter space while reducing computational cost. Table 4 summarizes the optimal settings for each model.

To ensure generalizability, cross-validation was applied using a Stratified K-Fold approach. This method maintains class distribution across folds and is particularly suited for imbalanced datasets. We used 10 folds, a commonly recommended value in the literature, to balance bias and variance and enable fair model comparison.

Table 4. Summary of machine learning models and their optimized hyperparameter settings the final settings selected via RandomizedSearchCV to balance performance and computational cost. Values correspond to the configurations used for the reported results

Model	Optimized Hyperparameters and Values
Decision Tree	criterion = entropy, max_depth = None, min_samples_split = 4, min_samples_leaf = 1
Random Forest	n_estimators = 108, max_depth = 20, max_features = sqrt, min_samples_split = 15, min_samples_leaf = 4, bootstrap = True
XGBoost	n_estimators = 444, max_depth = 6, learning_rate = 0.167, subsample = 0.646, colsample_bytree = 0.592, gamma = 0.152
SVM	C = 5.997, kernel = rbf, gamma = scale
KNN	n_neighbors = 1, weights = distance, metric = euclidean
Neural Network	hidden_layer_sizes = (100, 50), activation = "tanh", alpha = 0.00405, learning_rate = "adaptive", early_stopping = True, n_iter_no_change = 10, solver = "adam", validation_fraction = 0.1.

4. Results and Clinical Interpretability of AI Models for Thrombocytopenia Prediction

A total of six machine-learning algorithms were evaluated to predict the onset of thrombocytopenia in patients with atrial fibrillation. Performance was assessed via stratified 10-fold cross–validation to ensure robust estimates.

4.1. Model Performance Comparison:

We evaluated models using eight metrics: accuracy, precision, recall, F1-score, specificity, area under the receiver operating characteristic curve (AUC-ROC), Matthews correlation coefficient (MCC), and log-loss. Each metric was selected to capture a different dimension of model behavior, from overall correctness to error calibration and sensitivity to minority-class predictions.

A comprehensive comparison of model performance is presented in Table 5.

Table 5. Cross-validation performance of machine learning models across multiple metrics Summary of the accuracy, precision, recall, F1-score, specificity, AUC-ROC, MCC, and log-loss for six algorithms. XGBoost stands out overall, combining high discrimination with the lowest log-loss for better calibration

Model	Accuracy	F1_score	Precision	Recall	Specificity	AUC_ROC	MCC	Logloss
Decision Tree	97.92	94.37	96.17	94.44	100.0	96.23	93.97	243.16
Random Forest	92.04	76.28	100.0	63.7	97.3	96.89	75.48	31.44
XGBOOST	97.08	92.71	98.44	89.29	97.3	98.91	91.79	17.74
SVM	78.66	9.68	37.5	5.77	100.0	56.33	11.67	54.99
KNN	71.59	36.12	38.81	35.71	67.57	58.65	18.74	1100.37
Neural Network	77.83	9.13	25.71	16.99	97.3	54.39	9.23	61.13

XGBoost and Decision Tree emerged as the top performers, achieving the highest accuracy (97.08% and 97.92%), F1-score (92.71% and 94.37%), and AUC-ROC (98.91% and 96.23%). In contrast, SVM, KNN and the Neural Network struggled with this imbalanced classification task, yielding low F1-scores (9.68%, 36.12 %, 9.13%) and recall rates (5.77%, 35.71%, 16.99%).

Notably, XGBoost achieved the lowest log-loss (17.74), indicating superior probability calibration, while its Matthews correlation coefficient (91.79%) confirmed its robustness under class imbalance.

4.2. Holdout Test Performance:

To evaluate real-world generalizability, we assessed all models on a 20% holdout test set kept entirely separate from training and tuning. XGBoost achieved 93.75% accuracy, 85.71% F1-score, and the highest AUC-ROC (98.03%), indicating strong performance on unseen data. The Decision Tree obtained a slightly higher accuracy (97.92%) but a much worse log-loss (75.09%), reflecting poorer probability calibration than XGBoost (16.93%). Random Forest performed well overall (AUC = 92.63%) but remained below XGBoost. Taken together, these results show that XGBoost offers the best balance between discrimination and calibration, and therefore is selected as the final model for thrombocytopenia prediction.

Table 6. Confusion matrix results for all classifiers on the test set TN and TP represent correctly classified negative and positive cases, respectively, while FP and FN correspond to misclassifications. Ensemble models (XGBoost, Random Forest) achieved the best overall balance, minimizing both false positives and false negatives

Model	True Negatives (TN)	False Positives (FP)	False Negatives (FN)	True Positives (TP)
Decision Tree	37	0	1	10
Random Forest	36	1	3	8
XGBoost	36	1	2	9
SVM Classifier	37	0	9	2
KNN	25	12	9	2
Neural Network	13	24	7	4

On a moderately imbalanced dataset (22.2% positives), the test-set confusion matrices (Table 6) show that high accuracy can conceal clinically important false negatives. For example, SVM produced no false positives but missed nine positive cases. By contrast, Decision Tree (1 FN, 0 FP) and XGBoost (2 FN, 1 FP) achieve a more realistic balance between sensitivity and precision. These findings temper accuracies above 97% and suggest better generalizability for tree and ensemble models, although external validation would still be useful.

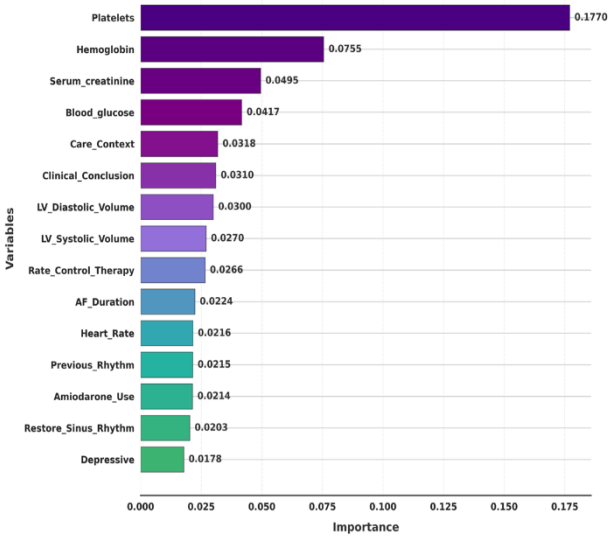
4.3. Feature Importance:

Among all features, platelet count emerged as the most influential predictor (score = 0.1770), reflecting its critical role in assessing bleeding risk, particularly in anticoagulated patients. Other major contributors included hemoglobin level (0.0755), creatinine level (0.0495), and blood glucose (0.0417), highlighting the interplay between hematologic status, renal function, and metabolic balance.

Structural cardiac parameters, namely, left ventricular diastolic and systolic volumes, also ranked highly, pointing to the importance of cardiac remodeling in atrial fibrillation. Clinical context variables such as care setting and diagnostic summary further informed the model, suggesting that care accessibility and physician judgment carry predictive value. Additionally, therapeutic variables, rate control strategies, use of amiodarone, and attempts to restore sinus rhythm, contributed meaningfully to outcome prediction.

Notably, depressive disorder, though lower in ranking (0.0100), remained a non-negligible factor, reinforcing the relevance of psychological well-being in patient trajectories. Together, these findings indicate that the model captures a multidimensional profile encompassing both physiological and psychosocial domains.

Figure 3: Top 15 Most Important Features According to Random Forest



4.4. Discrimination and Calibration Analysis:

We assessed how well the models separate cases and how trustworthy their probabilities are. Figure 4 reports Precision–Recall (PR) curves, which are more informative than ROC under our moderate imbalance. XGBoost and Random Forest keep precision high as recall increases, Decision Tree is close behind, while SVM, KNN, and Neural Network lose precision at higher recall.

Figure 4. Precision-recall curves of the machine learning models for thrombocytopenia prediction

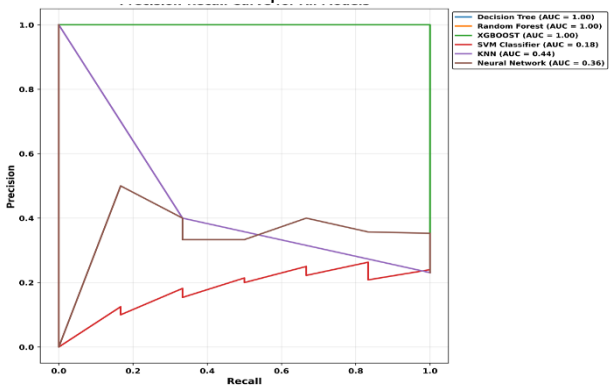
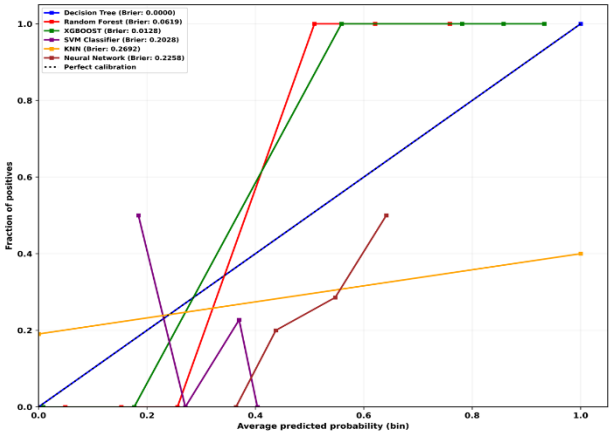


Figure 5 shows calibration plots, comparing predicted risks with observed event rates. XGBoost and Random Forest track the ideal diagonal closely, Decision Tree tends to overestimate risk, and SVM/Neural Network are less well-aligned.

Figure 5. Calibration Plots for all Machine Learning Models



Overall, ensemble models pair strong discrimination with reliable probabilities, making them the most suitable for thrombocytopenia prediction.

4.5. SHAP-Based Interpretability Analysis of the Thrombocytopenia Prediction Model:

The SHAP analysis (Figure 6) revealed that platelet count was the strongest contributor to model predictions. Low platelet levels were consistently associated with a higher predicted risk, likely due to their relevance in bleeding complications under anticoagulation therapy. Hemoglobin concentration followed closely, where lower values indicated increased patient vulnerability, reinforcing the known association between anemia and poor clinical outcomes.

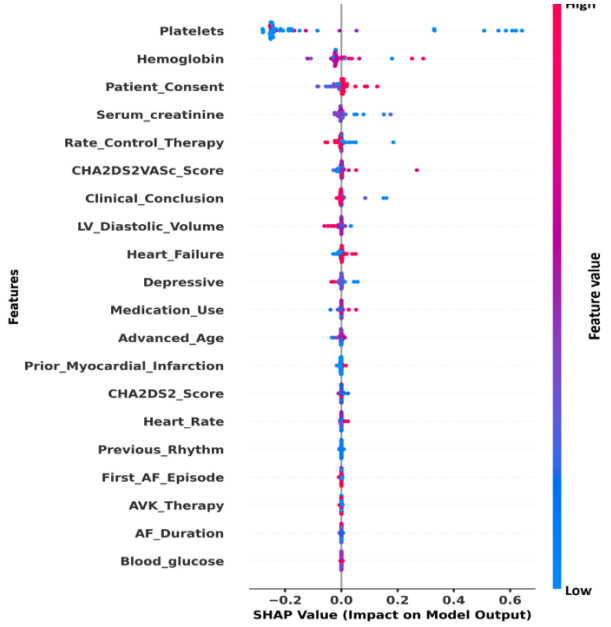
Interestingly, patient consent also emerged as a significant predictor, potentially capturing subtle variations in care engagement, documentation practices, or institutional processes. Serum creatinine levels, reflective of renal function, ranked among the top features as well, aligning with the well-established impact of kidney impairment on therapeutic decisions in atrial fibrillation.

Several therapeutic and structural variables also showed meaningful contributions. The use of rate control therapy, amiodarone, and attempts to restore sinus rhythm likely signaled more advanced or symptomatic arrhythmia profiles. Echocardiographic measurements such as left ventricular diastolic volume and the presence of heart failure further emphasized the importance of cardiac remodeling and dysfunction in the risk landscape.

Clinical scoring systems like the CHA_2DS_2-VASc and CHA_2DS_2 scores were appropriately ranked, confirming that the model's outputs are grounded in validated risk stratification frameworks. Lastly, the presence of depressive symptoms, while less dominant, still influenced predictions, underscoring the relevance of psychological health in overall patient prognosis.

Some non-clinical features (e.g., consent, depressive symptoms) likely reflect site or reporting bias rather than biology and should be interpreted with caution. We do not treat these variables as causal predictors, and they do not inform any clinical recommendations in this study. Collectively, the model integrates a broad spectrum of clinical, biological, therapeutic, and psychosocial variables, offering a multidimensional approach to risk prediction in atrial fibrillation.

Figure 6. SHAP Summary Plot Showing Feature Contributions to the Thrombocytopenia Prediction Model



4.6. Review of Model Findings and Recommendations:

The results of this study highlight the relevance of supervised learning for predicting thrombocytopenia in patients with atrial fibrillation. Among the models tested, XGBoost and Decision Tree achieved the highest performance, particularly in terms of accuracy, F1-score, AUC-ROC, and MCC, while also managing class imbalance effectively. XGBoost proved to be the most robust model overall, with strong calibration, a near-perfect AUC (98.91%), and an excellent trade-off between precision and recall. The Decision Tree also performed well, with the added benefit of interpretability.

SHAP analysis reinforced the clinical validity of XGBoost by identifying platelet count, hemoglobin, and creatinine as key predictive variables, consistent with known medical indicators. It also enabled interpretation of individual predictions, enhancing the model's transparency.

In summary, XGBoost is recommended for practical clinical use, while the Decision Tree offers a reliable, explainable alternative. Both approaches, strengthened by SHAP explainability, provide effective and interpretable tools for risk prediction in this context.

This section offers a critical interpretation of the study's findings, highlighting the contributions of the proposed approach, examining its limitations, and exploring the broader implications of using machine learning to predict thrombocytopenia in patients with atrial fibrillation.

5. Discussion

The results of this study shed valuable light on how artificial intelligence, particularly supervised machine learning, can contribute to improving risk prediction in patients with atrial fibrillation. In what follows, we explore the robustness and interpretability of the models developed, highlight the added value they bring beyond conventional clinical tools, critically examine the study's limitations, and suggest directions for future research and clinical application.

5.1. Model Strengths and Interpretability:

This study confirms the effectiveness of supervised learning models for predicting thrombocytopenia in atrial fibrillation (AF) patients using real-world data from sub-Saharan Africa. Among the tested classifiers, ensemble methods, especially XGBoost and Random Forest, stood out for their accuracy and interpretability, echoing prior findings on the reliability of tree-based and boosting techniques in structured clinical data (Chen and Guestrin, 2016). SHAP analysis revealed platelet count as the most influential variable, in line with the known biology of thrombocytopenia (Stasi, 2012). Other important features included hemoglobin, serum creatinine, use of vitamin K antagonists, and heart rate control (RCA). The model's ability to integrate this range of clinical parameters illustrates its superiority over traditional scores, which often depend on a narrow set of fixed variables (Rahul *et al.*, 2025). In addition to its strong predictive power, the model's SHAP-based interpretability enhances its clinical relevance by explaining how different features contribute to individual risk. This transparency is vital for building trust and ensuring integration into medical practice.

5.2. Added Value of Machine Learning Over Conventional Clinical Tools:

Unlike fixed clinical scores ($CHADS_2$, CHA_2DS_2-VASc, HAS-BLED), ML captures complex, non-linear interactions across many variables. This enables a more refined and individualized assessment of risk, particularly for multifactorial conditions like thrombocytopenia (Lu *et al.*, 2024). The integration of SHAP values further enhanced the interpretability of the models, translating technical outputs into a

format that is intuitive and clinically meaningful. This transparency is crucial for fostering trust and facilitating the integration of such models into everyday medical decision-making (Cheng *et al.*, 2021). The analysis revealed key predictors including platelet count, hemoglobin level, serum creatinine, rhythm control therapies, and the CHA_2DS_2-VASc score, findings that align well with established clinical understanding (Iyengar *et al.*, 2023). Notably, the model also recognized the importance of secondary variables such as age and treatment history, underscoring its ability to reflect the nuanced nature of patient risk profiles.

From a pathophysiological perspective, the SHAP top predictors are biologically coherent. Low platelet and hemoglobin levels indicate hematologic depletion, typical of thrombocytopenia and anemia that often coexist in AF. Higher creatinine points to renal impairment, which contributes to platelet dysfunction and bleeding risk (Iijima *et al.*, 2023). Age further magnifies these effects through accumulated comorbidities and frailty, supporting the biological plausibility of the model's predictions. Of particular interest is the influence of less prominent factors like prior rhythm abnormalities or certain medications, which, although more subtle, provide valuable insights into risk stratification. These findings may open new avenues for research into understudied clinical features and support the development of more precise diagnostic and prognostic approaches.

5.3. Limitations and Potential Biases:

While certain limitations must be acknowledged, they do not undermine the validity of the conclusions. Although the data were collected from multiple healthcare centers, the overall sample size (n = 239) remains relatively modest. With a modest sample (n = 239) and many predictors, some overfitting risk persists despite stratified cross-validation and SMOTE. Findings are context-bound to Côte d'Ivoire and need multi-country AFRICA validation to confirm applicability elsewhere. The imbalance between patients with and without thrombocytopenia represents a well-known challenge in this type of research. This issue was partially mitigated through stratified cross-validation and the use of the Matthews correlation coefficient (MCC), but it should still be kept in mind when interpreting the results. The predictive models, although encouraging, would benefit from further testing on larger datasets and in real-world clinical environments. This would help assess their ability to handle clinical variability and confirm their reliability over time (Rahul et al., 2025). That said, these limitations do not undermine the main findings. Rather, they point to the value of future studies to build on and strengthen the current results.

6. Conclusion

This study developed and evaluated a predictive model for thrombocytopenia among patients with atrial fibrillation (AF) using supervised machine learning. Based on a multicenter dataset from Côte d'Ivoire and rigorous validation, the results showed that tree-based algorithms, particularly XGBoost, achieved superior predictive performance and robust handling of class imbalance. Beyond accuracy, the integration of SHAP (SHapley Additive exPlanations) provided interpretability, revealing biologically consistent predictors such as platelet count and serum creatinine, alongside additional behavioral and treatment-related factors. These findings highlight the model's ability to integrate diverse clinical dimensions into coherent, data-driven insights. The results demonstrate that artificial intelligence can effectively identify AF patients at high risk of thrombocytopenia, supporting earlier and more targeted clinical intervention. However, given the modest sample size and the single-country origin of the data, external validation across broader African populations remains essential to confirm the model's generalizability and robustness. Looking forward, integrating this model into Clinical Decision Support Systems (CDSS) could strengthen risk assessment and guide proactive anticoagulation management. Such tools, when rigorously validated and transparently deployed, can help bridge the gap between predictive analytics and real-world clinical care in Africa. In conclusion, this study illustrates the growing potential of explainable machine learning in medicine, particularly in the management of complex conditions such as atrial fibrillation. When carefully implemented and transparently interpreted, such technologies can provide valuable data-driven decision support in a healthcare environment where anticipating risks is becoming increasingly essential.

Data Availability Statement

The original clinical data are not publicly available for confidentiality reasons. To support reproducibility, a synthetic dataset with the same structure and the full analysis code are on GitHub:

https://github.com/AmadouDiabagate/Thrombocytopenia_Prediction_ML. Access to anonymized clinical data may be considered upon reasonable request and required approvals.

Acknowledgement

The authors gratefully acknowledge the support provided by Faculty of Mathematics and Computer Science, University Felix Houphouët-Boigny.

Funding

This research did not receive a specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare no conflict of interest.

Biographies

Amadou Diabagate

Faculty of Mathematics and Computer Science, University Felix Houphouët-Boigny, Cocody, Abidjan, Côte d'Ivoire, 002250767954216, amadou1.diabagate@ufhb.edu.ci

Amadou earned his Ph.D. in Computer Science (Artificial Intelligence) from the Faculty of Science and Technology, Abdelmalek Essaadi University, Tangier, Morocco, in 2016. Since 2018, he has been an Assistant Professor at the Faculty of Mathematics and Computer Science, University Felix Houphouët-Boigny, Côte d'Ivoire. His research interests include artificial intelligence, medical data science, and digital transformation. He has published in several indexed journals and actively contributes to national digital transformation initiatives.

ORCID: 0000-0002-8363-0871

Katienefowa Sekou Koulibaly

Faculty of Medicine, University Felix Houphouët-Boigny, Abidjan, Côte d'Ivoire, 002250747496603, sekou.koulibaly@mugef-ci.com

Katienefowa earned his medical doctorate from Félix Houphouët-Boigny University in Côte d'Ivoire. He has extensive experience in managing patients with atrial fibrillation and currently serves as a medical advisor at the Mutuelle Générale des Fonctionnaires de Côte d'Ivoire (MUGEF-CI). Alongside his professional duties, he is pursuing specialized training in Radiology and Medical Imaging. Driven by a strong commitment to continuous professional development and to advancing healthcare efficiency, Dr. Koulibaly is particularly interested in atrial fibrillation and in modernizing clinical practice.

ORCID: 0009-0007-6306-4625

Awa Fofana

Faculty of Medicine, University Felix Houphouët-Boigny, Abidjan, Côte d'Ivoire, 002250777148036, yeninfofana1995@gmail.com

Awa earned her Doctor of Medicine degree in 2022 from the Faculty of Medicine at Félix Houphouët-Boigny University in Abidjan, Côte d'Ivoire. She currently serves as Head of the Mother and Child Unit at Nassian General Hospital. While pursuing a specialization in medical oncology at the same faculty, she has gained clinical experience in hematology, particularly in the management of patients with blood cancers. In addition to her clinical work, she actively participates in health-promotion activities and community outreach initiatives.

ORCID: 0009-0008-2891-1697

Doffou Jérôme Diako

Ecole Supérieure Africaine des TIC (ESATIC), Abidjan, Côte d'Ivoire, 002250102046342, jerome.diako@esatic.edu.ci

Diako holds a Ph.D. in Cyberdefense and Artificial Intelligence and currently serves as an Assistant Professor at ESATIC (École Supérieure Africaine des Technologies de l'Information et de la Communication) in Abidjan, Côte d'Ivoire. His areas of expertise include cybersecurity, artificial intelligence, and digital transformation. He coordinates graduate programs in Big Data, Cybersecurity, and AI, and actively contributes to national initiatives on digital governance and cyber resilience. He has published in indexed scientific journals on intelligent threat detection, machine learning applications, and AI ethics in African contexts.

ORCID: 0000-0002-8815-587X

References

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Breiman, L. (2001). Random forests. *Machine Learning,* **45**(1), 5–32. DOI: 10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R.A. and Stone, C.J. (2017). Classification and *Regression Trees*. *Chapman and Hall/CRC*. DOI: 10.1201/9781315139470

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). DOI: 10.1145/2939672.2939785

Cheng, Y., Chen, C., Yang, J., Yang, H., Fu, M., Zhong, X., Wang, B., He, M., Hu, Z., Zhang, Z., Jin, X., Kang, Y. and Wu, Q. (2021). Using machine learning algorithms to predict hospital-acquired thrombocytopenia after operation in the intensive care unit: A retrospective cohort study. *Diagnostics,* **11**(9), 1614. DOI: 10.3390/diagnostics11091614

Chugh, S.S., Havmoeller, R., Narayanan, K., Singh, D., Rienstra, M., Benjamin, E.J., Gillum, R.F., Kim, Y.-H., McAnulty, J.H., Zheng, Z.-J., Forouzanfar, M.H., Naghavi, M., Mensah, G.A., Ezzati, M. and Murray, C.J.L. (2014). Worldwide epidemiology of atrial fibrillation: A global burden of disease 2010 study. *Circulation,* **129**(8), 837–47. DOI: 10.1161/CIRCULATIONAHA.113.005119

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning,* **20**(3), 73–297. DOI: 10.1007/BF00994018

Ding, L. and Tang, M. (2025). Global, regional and national burden of atrial fibrillation and flutter attributable to metabolic risks from 1990 to 2021: Analysis of data from the global burden of disease study. *EP Europace,* **27**(1), euaf085.269. DOI: 10.1093/europace/euaf085.269

Diop, K.R., Samb, C.A.B., Kane, A., Mingou, J.S., Beye, S.M., Diouf, Y. and Adoubi, A. K. (2022). Atrial fibrillation in three cardiological reference centers in Dakar: Senegal data from the AFRICA register survey. *The Pan African Medical Journal*, **43**(n/a), 112. DOI: 10.11604/pamj.2022.43.112.31397

Ettarh, R. (2016). Patterns of international collaboration in cardiovascular research in sub-saharan Africa. *Cardiovascular Journal of Africa,* **27**(3), 436–45. DOI: 10.5830/CVJA-2015-082

Ghanima, W. and Cooper, N. (2024). Could machine learning revolutionize how we treat immune thrombocytopenia? *British Journal of Haematology*, **205**(3), 770–1. DOI: 10.1111/bjh.19684

Haibo, H. and Garcia, E.A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering,* **21**(9), 1263–84. DOI: 10.1109/TKDE.2008.239

Haroon, M., Dimitrios, K., Giacinto, L.P., Drew, P. and Frederick, C. (2024). Can machine learning assist in diagnosis of primary immune thrombocytopenia? A feasibility study. *Diagnostics,* **14**(13), 1352. DOI: 10.3390/diagnostics14131352

Hindricks, G., Potpara, T., Dagres, N., Arbelo, E., Bax, J.J., Blomström-Lundqvist, C., Boriani, G., Castella, M., Dan, G.A., Dilaveris, P.E., Fauchier, L., Filippatos, G., Kalman, J.M., La Meir, M., Lane, D.A., Lebeau, J.P., Lettino, M., Lip, G.Y.H., Pinto, F.J., Thomas, G.N., Valgimigli, M., Van Gelder, I.C., Van Putte, B.P. and Watkins, C.L. (2021). 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation. *European Heart Journal,* **42**(5), 373–498. DOI: 10.1093/eurheartj/ehaa612

Iijima, R., Tokue, M., Nakamura, M., Yasuda, S., Kaikita, K., Akao, M., Ako, J., Matoba, T., Miyauchi, K., Hagiwara, N., Kimura, K., Hirayama, A., Matsui, K., Ogawa, H. and AFIRE Investigators. (2023). Thrombocytopenia as a bleeding risk factor in atrial fibrillation and coronary artery disease: Insights from the AFIRE study. *Journal of the American Heart Association,* **12**(20), e031096. DOI: 10.1161/JAHA.123.031096

Iyengar, V., Patell, R., Ren, S., Ma, S., Pinson, A., Barnett, A., Elavalakanar, P., Kazi, D.S., Neuberg, D. and Zwicker, J.I. (2023). Influence of thrombocytopenia on bleeding and vascular events in atrial fibrillation. *Blood Advances,* **7**(24), 7516–24. DOI: 10.1182/bloodadvances.2023011235

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R* (Vol. 103). New York: Springer. DOI: 10.1007/978-1-4614-7138-7

January, C.T., Wann, L.S., Calkins, H., Chen, L.Y., Cigarroa, J.E., Cleveland, J.C., Ellinor, P.T., Ezekowitz, M.D., Field, M.E., Furie, K.L., ... (2019). AHA/ACC/HRS focused update on atrial fibrillation: Management guidelines. *Circulation,* **140**(2), e125–e151. DOI: 10.1161/CIR.0000000000000665

Jiang, X., Wang, Y., Pan, Y. and Zhang, W. (2022). Prediction models for sepsis-associated thrombocytopenia risk in intensive care units based on a machine learning algorithm. *Frontiers in Medicine*, **9**(n/a), 837382. DOI: 10.3389/fmed.2022.837382

Lip, G., Gue, Y., Zhang, J., Chao, T., Calkins, H. and Potpara, T. (2022). Stroke prevention in atrial fibrillation. *Trends in Cardiovascular Medicine,* **32**(8), 501–10. DOI: 10.1016/j.tcm.2021.10.001

Liu, T., Krentz, A., Lu, L. and Curcin, V. (2025). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis. *European Heart Journal-Digital Health*, **6**(1), 7–22. DOI: 10.1093/ehjdh/ztae080

Lu, J., Bisson, A., Bennamoun, M., Zheng, Y., Sanfilippo, F.M., Hung, J., Briffa, T., McQuillan, B., Stewart, J., Figtree, G., Huisman, M.V., Dwivedi, G. and Lip, G.Y. (2024). Predicting multifaceted risks using machine learning in atrial fibrillation: Insights from GLORIA-AF study. *European Heart Journal – Digital Health*, **5**(3), 235–46. DOI: 10.1093/ehjdh/ztae010

Lundberg, S.M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems,* **30**(n/a), 4765–74. DOI: 10.48550/arXiv.1705.07874

Mendis, S., Puska, P. and Norrving, B. (2011). *Global Atlas on Cardiovascular Disease Prevention and Control*. World Health Organization.

Miotto, R., Wang, F., Wang, S., Jiang, X. and Dudley, J.T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics,* **19**(6), 1236–46. DOI: 10.1093/bib/bbx044

Moulis, G., Comont, T., Germain, J., Sommet, A., Lapeyre-Mestre, M., Beyne-Rauzy, O. and Adoue, D. (2020). Significance of antinuclear antibodies in primary immune thrombocytopenia: Results of the CARMEN registry. *Blood Advances,* **4**(9), 1974–7. DOI: 10.1182/bloodadvances.2020001664

Pisters, R., Lane, D.A., Nieuwlaat, R., de Vos, C.B., Crijns, H.J.G.M. and Lip, G.Y.H. (2010). The HAS-BLED score: A novel tool to assess bleeding risk in patients with atrial fibrillation. *Chest,* **138**(5), 1093–100. DOI: 10.1378/chest.10-0134

Rahul, C., Mehdi, N., Floyd, W.T., Walid, F.G., Wei-Hsuan, L.-C., Rohit, C., Anahita, D., Kevin, P.B., Paul, A.G., Matthew, D.N., Sandeep, J., Aditya, B., Suresh, R.M., Yanshan, W., Matthew, E.H. and Samir, S. (2025). Machine learning predicts bleeding risk in atrial fibrillation patients on direct oral anticoagulant. *The American Journal of Cardiology,* **244**(n/a), 58–66. DOI: 10.1016/j.amjcard.2025.02.030

Rajkomar, A., Dean, J. and Kohane, I. (2019). Machine learning in medicine. *The New England Journal of Medicine,* **380**(14), 1347–58. DOI: 10.1056/NEJMra1814259

Stambler, B.S. and Ngunga, L.M. (2015). Atrial fibrillation in sub-Saharan Africa: Epidemiology, unmet needs and treatment options. *International Journal of General Medicine,* **8**(n/a), 231–42. DOI: 10.2147/IJGM.S84537

Stasi, R. (2012). Immune thrombocytopenia: Pathophysiologic and clinical update. *Seminars in Thrombosis and Hemostasis,* **38**(5), 454–62. DOI: 10.1055/s-0032-1305780

Verheugt, F.W. and Granger, C.B. (2015). Oral anticoagulants for stroke prevention in atrial fibrillation: Current status, special situations and unmet needs. *The Lancet,* **386**(9990), 303–10. DOI: 10.1016/S0140-6736(15)60245-8

Yeh, Y.H., Chan, Y.H., Chen, S.W., Chang, S.H., Wang, C.L., Kuo, C.T. and Chao, T.F. (2022). Oral anticoagulant use for patients with atrial fibrillation with concomitant anemia and/or thrombocytopenia. *The American Journal of Medicine*, **135**(8), e248-56. DOI: 10.1016/j.amjmed.2022.03.011

Copyright