npj Digital Medicine

https://doi.org/10.1038/s41746-026-02436-5

Article in Press

A deep learning model integrating structured data and clinical text for predicting atrial fibrillation recurrence

Received: 16 September 2025

Accepted: 5 February 2026

Cite this article as: Jia, S., Yin, Y., Guan, Y. *et al.* A deep learning model integrating structured data and clinical text for predicting atrial fibrillation recurrence. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02436-5

Sixiang Jia, Yanping Yin, Yingxia Guan, Xuan Ying, Peijian Shi, Chenhao Li, Qiang Yang, Xuanting Mou, Jiangbo Lin, Que Xu, Qingru Zhu, Yang Yang, Heqiang Zhang, Jianqiang Zhao, Wenting Lin, Chao Feng, Weili Ge & Shudong Xia

A deep learning model integrating structured data and clinical text for predicting atrial fibrillation recurrence

Sixiang Jia [1,#,*], Yanping Yin [2,#,*], Yingxia Guan [3,#], Xuan Ying [4,#], Peijian Shi [5], Chenhao Li[1], Qiang Yang [6], Xuanting Mou [7], Jiangbo Lin [2], Que Xu [4], Qingru Zhu [1], Yang Yang [1], Heqiang Zhang[1], Jianqiang Zhao[1], Wenting Lin [1], Chao Feng [1], Weili Ge [2,*], Shudong Xia [1,*]

1. Department of Cardiology, The Fourth Affiliated Hospital, Zhejiang University School of Medicine, Yiwu, Zhejiang 322000, China
2. Department of Cardiology, Taizhou Hospital of Zhejiang Province, Wenzhou Medical University, Linhai, Zhejiang 317000, China
3. Department of Cardiology, Affiliated Hospital of Yunnan University, Kunming, Yunnan 650021, China
4. Department of Cardiology, Jinhua People's Hospital, Jinhua, Zhejiang 321000, China
5. Department of Cardiology, Beilun District People's Hospital, Ningbo, Zhejiang 315800, China
6. Department of Cardiology, The Second Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, Zhejiang 310000, China
7. Department of Cardiology, Taizhou First People's Hospital, Taizhou, Zhejiang 318020, China
# These authors contributed equally to this article

* Correspondence:

Shudong Xia, Department of Cardiology, The Fourth Affiliated Hospital, Zhejiang University School of Medicine, Yiwu, Zhejiang 322000, China
Emai:shystone@zju.edu.cn

Sixiang Jia, Department of Cardiology, The Fourth Affiliated Hospital, Zhejiang University School of Medicine, Yiwu, Zhejiang 322000, China
Emai:jiasixiang@zju.edu.cn

Weili Ge, Department of Cardiology, Taizhou Hospital of Zhejiang Province, Wenzhou Medical University, Linhai, Zhejiang 317000, China
Email: gewl@enzemed.com

Yanping Yin, Department of Cardiology, Taizhou Hospital of Zhejiang Province, Wenzhou Medical University, Linhai, Zhejiang 317000, China
Email: yinyp7713@enzemed.com

Abstract

Multimodal perioperative data from patients undergoing atrial fibrillation (AF) ablation offer valuable insights for stratifying recurrence risk, yet remain underutilized in prediction models. This multicenter retrospective study included 2,508 patients who underwent AF ablation at five Chinese centers: The Fourth Affiliated Hospital of Zhejiang University School of Medicine (Jan 2016–Mar 2024; Training Cohort), Taizhou Hospital of Zhejiang Province (Jan 2015–Jan 2024; Training Cohort), The Affiliated Hospital of Yunnan University (Jan 2016–Jan 2024; Validation Cohort), Jinhua People's Hospital (Jan 2020–Jan 2024; Test Cohort), and Ningbo Beilun Hospital (Jan 2020–Jan 2024; Test Cohort). We developed a dual-branch deep learning model to predict AF recurrence, in which structured data were processed via a 1D ResNet and textual data were encoded using four large language models (LLaMA-7B, Phi2-2.7B, Mistral-7B, and MedGemma-27B). The model incorporating MedGemma for text feature extraction performed best, achieving areas under the curve of 0.934 (95% CI: 0.921 – 0.946), 0.928 (95% CI: 0.904 – 0.950), and 0.911 (95% CI: 0.878 – 0.941) on the training, validation, and test sets, respectively. Our model integrates multimodal perioperative data from AF ablation patients, effectively identifies high-risk individuals, and may facilitate targeted interventions to reduce relapse.

Keywords: atrial fibrillation recurrence; catheter ablation; large language models; deep learning; multimodal medical text

Introduction

Catheter ablation (CA) is a rhythm control strategy endorsed by guidelines for patients with atrial fibrillation (AF) [1]. Nonetheless, long-term freedom from atrial arrhythmias remains unsatisfactory, with 30%–50% of patients experiencing recurrence within one year and requiring repeat ablation [2]. The identification of populations at high risk for AF recurrence facilitates the development of post-ablation preventive strategies and rhythm control regimens [2].

Established predictors, including AF type, left atrial diameter, AF duration, machine learning models based on structured preoperative data, and clinical risk scores such as APPLE and CAAP-AF, have demonstrated efficacy in forecasting recurrence [3,4,5]. However, they remain constrained by their reliance on pre-procedural information and fail to capture the impact of individualized treatment strategies and procedural specifics on outcomes. The clinical management of AF generates extensive multimodal textual records that encapsulate valuable information beyond the structured data. Echocardiography reports provide quantitative and qualitative assessments of the cardiac structure and function, incorporating interpreter-specific insights. Pre-procedural 24-hour Holter monitoring captures the dynamic arrhythmia burden and heart rate variability. Crucially, electrophysiologists document detailed procedural notes, including the ablation strategy, lesion set design, and real-time parameters, which are often lost in structured data systems. The semantic depth of these texts represents a critical and underutilized source of prognostic information.

Large language models (LLMs) provide a transformative approach to clinical text mining through advanced semantic understanding [6]. This study innovatively applied four state-of-the-art LLMs—MedGemma, Phi-2, Llama, and Mistral—to derive high-dimensional features from clinical texts such as echocardiography reports, Holter interpretations, and ablation notes [7,8,9,10]. By evaluating the proficiency of each model in medical language comprehension, we identified the most predictive LLM for textual representation. We further propose a deep learning framework that combines LLM-based text features with structured clinical data. Using peri-procedural text extracted from hospital systems for patients with AF undergoing ablation, we aimed to develop a precise and individualized AF recurrence prediction model tailored to personalized treatment.

Results

Participant Baseline

This study retrospectively analyzed data from patients who underwent AF ablation at five Chinese AF centers: the Fourth Affiliated Hospital of Zhejiang University School of Medicine (ZJU4th; January 2016–March 2024), Taizhou Hospital of Zhejiang Province (ZJTZH; January 2015–January 2024), the Affiliated Hospital of Yunnan University (YNH; January 2016–January 2024), Jinhua People's Hospital (JHPH; January 2020–January 2024), and Ningbo Beilun Hospital (NBH; January 2020–January 2024). As shown in Table 1 (the missing structured data is shown in **Supplementary Table 1**), a total of 2,508 participants were enrolled in this study, with a median age of 65.00 (interquartile range [IQR] 58.00–71.00) years. Among them, 1,572 (62.68%) were

men. The overall median follow-up duration was 31.00 (IQR 17.00–48.00) months, and the overall recurrence rate was 22.57%. Significant differences ($p < 0.05$) were observed across centers in the following variables: age, systolic blood pressure (SBP), diastolic blood pressure (DBP), AF duration, CHA_2DS_2-VASc score, HAS-BLED score, left atrial diameter (LAD), left ventricular ejection fraction (LVEF), survival time, high-density lipoprotein (HDL), low-density lipoprotein (LDL), albumin, creatinine, estimated glomerular filtration rate (eGFR), APPLE score, CAAP-AF score, gender, hypertension, coronary artery disease, diabetes, AF type, and use of class I/III or class II antiarrhythmic drugs.

Model Development and Validation

Before model development, patient data from the five centers were assigned to a training set (ZJU4th and ZJTZH), a validation set (YNH), and a test set (JHPH and NBH). Detailed data distributions are provided in **Supplementary Table 2, Supplementary Table 3, Supplementary Table 4**. Within the architecture of the dual-branch deep learning network, we fine-tuned different LLMs (LLaMA-7B, Phi2-2.7B, Mistral-7B, and MedGemma-27B) as the base for the textual feature extraction branch (The clinical text data are provided in the **Supplementary 3**), while keeping the structured data branch and late fusion pathway identical across experiments. As shown in **Figure 1A–C**, the model incorporating the MedGemma-27B module for text processing, followed by late fusion with structured data (MedGemma-Fusion), yielded the best performance. It achieved an area under the curve of 0.934 (95% confidence interval [CI]: 0.921–0.946) in the training set, 0.928 (95% CI: 0.904–0.950) in the validation set, and 0.911 (95% CI: 0.878–0.941) in the test set. Additionally, we assessed the impact of sample size on the performance of the MedGemma-Fusion model (**Supplementary Figure 1)**. To evaluate the generalization performance of the model, we used each center as the test set, with the remaining centers constituting the training and validation sets (**Supplementary Table 5)**. To further validate the rationale for late fusion, we conducted additional ablation experiments, including predictions of AF recurrence based solely on structured data and predictions based on text features extracted using the optimal LLM (MedGemma) (**Supplementary Table 6, Supplementary Figure 2**). Moreover, we conducted a series of ablation studies to rigorously substantiate the design of MedGemma-Fusion network for predicting AF recurrence. We adopted a two-stage fine-tuning strategy for the text feature extraction branch: domain-adaptive pre-training, followed by supervised contrastive fine-tuning (**Supplementary Table 7, Supplementary Figure 3**), during this process, the weights of the structured data channel in MedGemma-Fusion were frozen. In the structured data-processing branch, given that recurrence represents a relatively minor class, we implemented a conditional tabular generative adversarial network (GAN) within the training fold to perform data augmentation and balance the class distribution. Specifically, a conditional Wasserstein GAN with gradient penalty (WGAN-GP) variant tailored to tabular data was employed, where the generator was conditioned on class labels and encoded categorical features to produce synthetic positive samples consistent with the real joint distribution. After freezing the weights of the text feature channel in the optimally trained MedGemma-Fusion model, we compared different data augmentation techniques (synthetic minority oversampling technique [SMOTE] and adaptive synthetic sampling [ADASYN]) to demonstrate the advantages of using

the WGAN-GP (**Supplementary Table 8, Supplementary Figure 4**). We further compared the MedGemma-Fusion model against conventional AF recurrence risk factors (including AF type and LAD) and established clinical risk scores (CHA_2DS_2-VASc, CAAP-AF, and APPLE). As summarized in **Table 2** and depicted in **Figure 1D–L**, MedGemma-Fusion outperformed all reference models across all three datasets. Moreover, decision curve analysis (DCA) consistently demonstrated the superior clinical utility of MedGemma-Fusion across datasets, as shown in **Figure 1G–I**. We also benchmarked our novel AF recurrence risk model against models derived from the original variables of the APPLE and CAAP-AF scores (**Supplementary Table 9, Supplementary Table 10, Supplementary Table 11**).

Kaplan–Meier (K-M) Survival Analysis

Based on the optimal MedGemma-Fusion model, we generated K-M survival curves to evaluate its ability to discriminate recurrence risk across different datasets. As shown in **Figure 2**, the model exhibited a strong discriminative ability for recurrence risk, with a concordance index of 0.874 in the training set, 0.860 in the validation set, and 0.870 in the test set.

SHapley Additive exPlanations (SHAP)-Based Interpretability Analysis of the Model

In **Figure 3**, AF duration, LAD, and AF type emerged as the primary weighted variables within the structured data, whereas text-derived features such as "pulmonary vein," "potential," and "motion" were identified as key vectors contributing to the model's decision-making process.

Sensitivity Analysis

A sensitivity analysis was performed to evaluate the stability of the model with and without early AF recurrence (**Figure 4**).

Discussion

We developed a dual-branch deep learning network that integrates unstructured medical texts and structured data features based on perioperative data from patients undergoing AF ablation. In the text feature extraction module, a comparative evaluation of the four LLMs within an invariant fusion framework identified MedGemma-Fusion as optimal. This demonstrated robust predictive performance and generalizability in external validation.

In clinical practice, patients undergoing AF ablation generate multimodal medical data during the perioperative period, including unstructured medical texts and structured data. Textual modalities capture physicians' clinical interpretations and reflect patients' current AF progression. However, conventional machine learning models remain limited in their ability to comprehend the underlying logical relationships within medical services, which was corroborated by the models we built by employing clinical variables from the APPLE and CAAP-AF scores. In contrast, LLMs trained through extensive parameter iterations demonstrate superior performance in processing textual data. This study designed a dual-channel deep learning architecture that prevents structured data vectors from inducing hallucinatory interference during textual feature extraction. Furthermore, to fine-tune the LLMs, we employed a weakly supervised training strategy (recurrence risk classification) to mitigate catastrophic forgetting during text feature learning. The rationality of our network architecture was verified through a series of ablation studies.

The MedGemma-Fusion framework achieved the best predictive performance, indicating that

MedGemma effectively learned more discriminative feature representations from the multimodal medical text data of patients with AF. MedGemma, developed by Google (Mountain View, CA, USA) based on the Gemma-3 architecture, is a domain-specific LLM optimized for multimodal medical comprehension [7]. Its medical background knowledge enhances its utility in supporting clinical decision-making. Therefore, MedGemma performed optimally in this study, given that its pre-training corpus was derived from medical domains. Because LLMs operate through semantic segmentation, we conducted a keyword analysis on the best-performing model to identify semantically salient features [11]. Terms such as "pulmonary vein" and "potential," which are key concepts documented during ablation procedures, emerged as critical tokens. This aligns with the semantic compression mechanism of transformer-based models, where attention mechanisms and subword tokenization condense core semantic information into highly informative tokens, such as the aggregated representation of "pulmonary vein" or the verbal center of "ablation," causing attribution to focus on these tokens. These terms correspond to essential steps in electrophysiological procedures, including pulmonary vein isolation, which is the cornerstone of AF ablation, and the elimination of fractionated or additional potentials to disrupt rotor formation and terminate re-entrant circuits [12]. This confirms that the model successfully captured clinically relevant feature vectors.

The token "Vein" reflects the necessity of femoral vein puncture for catheter insertion during AF ablation, increasing its frequency in clinical narratives. Furthermore, the token "Motion" likely originates from the model's interpretation of preoperative echocardiography reports and carries significant weight in feature importance. Different types of AF exhibit distinct atrial motion patterns, potentially related to atrial fibrosis [13,14]. Although echocardiography is operator-dependent, we mitigated this heterogeneity by incorporating multicenter datasets, thereby enhancing the robustness of the model. By leveraging textual reports from echocardiograms, we input functional and kinematic descriptors of cardiac chambers based on the sonographers' expertise, circumventing the variability in imaging parameters or machine differences. Interestingly, tokens from Holter reports contributed relatively little to the model likely because paroxysmal AF, which was the predominant subtype in our cohort, often resulted in normal Holter findings. While previous studies by Krasteva and Zhang highlighted the predictive value of Holter monitoring for AF detection, its limited influence may be attributable to the loss of granular electrographic features in textual representations [15,16].

In the structured data analysis, factors such as the duration of AF, AF type, and left atrial diameter were identified as the most significant predictors of AF recurrence, which was consistent with our previous findings [17]. These factors contribute to deterioration of the atrial substrate to varying degrees, thereby perpetuating arrhythmia [14,18]. These results validate the efficacy of our approach, which combines sample augmentation with a ResNet-based architecture within a structured data pipeline.

In this study, the rationale behind adopting a center-based split into training/validation/test sets was to ensure, without introducing data leakage, sufficient sample size and relatively balanced class distribution during the training phase. This approach enables stable and reproducible training and hyperparameter tuning for both our model and all comparator baselines, while also allowing a

genuine assessment of cross-center generalization during testing. The ZJU4th and ZJTZ cohorts were used for training: together they provide a larger overall sample size with a more balanced ratio of positive to negative cases. Their combination offers richer learning signals for the model and baseline methods, thereby reducing instability and randomness caused by insufficient training data or severe class imbalance. The YNH cohort served for external validation (hyperparameter tuning and model selection): an independent center not involved in training was reserved for early stopping, hyperparameter selection, and final model determination. This prevents "implicit tuning" from repeated experimentation on the test set and ensures objectivity in the evaluation pipeline. The JHPH and NBH cohorts were held out for external testing. The remaining centers were entirely reserved for final performance assessment to evaluate cross-center generalization. In particular, JHPH, a center with fewer positive samples and a skewed distribution, is more suitable as an external test set representing a challenging "real-world" scenario. Including it in training could introduce severe class imbalance, compromise training stability, and hinder a conservative and credible estimation of generalization performance. To further verify the robustness of our deep learning model, we also performed additional leave-one-center-out validation.

Our study had some inherent limitations. First, although the model was validated across multiple centers, the sample size remained limited, and further validation with larger cohorts is warranted. Second, while we incorporated multimodal medical text data, including echocardiography, Holter electrocardiography (ECG), and procedural records, variability in reporting styles and levels of detail among physicians may have introduced heterogeneity into the text-based features. Although personalized ablation strategies (e.g., intraoperative energy settings and ablation sites) have been tailored to patients with AF across different centers, their impact on feature extraction using LLMs from templated procedural notes remains elusive. In the structured data domain, we incorporated both routine pre-operative laboratory tests and basic patient information. While exporting these structured data from the Hospital Information System reflects real-world clinical practice, it inevitably introduces variables unrelated to AF recurrence that may influence the model decision-making. Moreover, it is unavoidable that certain risk factors relevant to AF recurrence are omitted due to their unavailability in current real-world settings or within the Hospital Information System. Finally, the LLMs used in this study were based on the most recent versions available at the time of research; however, given the rapid iteration of LLMs, the impact of future updates on model performance remains uncertain. Additionally, the substantial computational resources required to deploy such models may impede their implementation in resource-limited healthcare settings.

We developed a dual-branch deep learning network that integrates feature representations extracted from medical texts using MedGemma, a specialized LLM, with structured data features derived from the perioperative records of AF ablation patients. This integrated approach provides novel clinical insights into risk prognosis and enhances strategies for the post-procedural management of AF ablation.

Methods

This study retrospectively analyzed data from consecutively enrolled patients who underwent AF ablation at five Chinese AF centers: ZJU4th (January 2016–March 2024), ZJTZH (January 2015–

January 2024), YNH (January 2016–January 2024), JHPH (January 2020–January 2024), and NBH (January 2020–January 2024). All patients were followed up until March 2025, with a minimum follow-up of one year for each individual. The study was conducted under the Declaration of Helsinki and received approval from the leading Ethics Committee of the Fourth Affiliated Hospital of Zhejiang University School of Medicine (No. K2025068); due to the retrospective nature and full anonymization of imaging data, informed consent was waived.

Structured perioperative data of the patients undergoing AF were extracted from each center's medical record system, including basic demographic indicators (age, gender, body mass index), patients' preoperative vital signs (preoperative SBP and DBP), comorbidity profile of patients (hypertension, coronary artery disease, and diabetes), cardiac structure and function in patients (LAD, LVEF), clinical status of patients (AF duration, AF type, CHA_2DS_2-VASc score, HAS-BLED score, and use of class I/III or class II antiarrhythmic drugs at admission), and preoperative laboratory parameters (glycosylated hemoglobin, fasting plasma glucose, total cholesterol, triglycerides, HDL, LDL, albumin, alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), creatine, and eGFR). The APPLE and CAAP-AF scores were also calculated. Textual data included Holter ECG reports, preoperative echocardiography reports, and surgical records.

The study exclusion criteria were as follows:

1.Repeat ablation procedures;
2.Patients with valvular AF;
3.AF patients with New York Heart Association class IV heart failure;
4.Missing Holter ECG or preoperative echocardiography data;
5.Loss to follow-up after ablation.

The patient enrollment flowchart is shown in **Figure 5A**.

Definition of AF Recurrence

According to the latest American College of Cardiology/American Heart Association guidelines, AF recurrence is defined as the presence of atrial arrhythmias (atrial tachycardia, atrial flutter, or atrial fibrillation) lasting > 30 s after the procedure. In this study, recurrences occurring within the first 3 months postoperatively were classified as early recurrences, whereas those occurring after 3 months were defined as late recurrences [1].

Post-ablation Follow-up of AF

Following ablation, all patients underwent outpatient follow-up and 24-hour ambulatory electrocardiogram monitoring at 1, 3, and 6 months after the procedure. At 12 months after the procedure, outpatient follow-up and 7-day long-term ambulatory electrocardiogram monitoring were performed. Subsequently, outpatient follow-up and 24-hour ambulatory electrocardiogram monitoring were performed every 6 months.

Development of a Multimodal Deep Learning Network

Building upon multimodal fusion and interpretable learning, this study adapted and extended methods for predicting AF recurrence. It specifically compared the impact of four open-source LLMs (LLaMA-7B, Phi2-2.7B, Mistral-7B, and MedGemma-27B) on the representation of Holter ECG reports, echocardiography reports, and surgical records. A convolutional neural network was

employed in the structured feature branch for representation learning and classification. Furthermore, a GAN was introduced to augment the categories and mitigate the imbalance caused by the scarcity of recurrence samples. The dataset comprised multimodal information from the perioperative period and follow-up, including 28 structured features and textual data from Holter ECG reports, echocardiography reports, and surgical records.

Preprocessing started with systematic data cleaning on structured channels. For continuous variables, a combined outlier detection method based on clinically plausible range constraints and the IQR rule was used, with extreme outliers beyond the threshold truncated at quantiles while preserving order information. Missing values were handled using a multiple imputation strategy; continuous variables were predicted and imputed using regression models constructed with multiple imputation chained equations, with mean and variance adjustments to avoid shrinkage; and categorical variables were imputed using mode or conditional sampling under Bayesian smoothed frequency encoding to preserve category co-occurrence relationships. To standardize the scales, continuous features were z-score normalized while retaining scaling parameters for external validation, and categorical variables were subjected to target leakage-free one-hot encoding or ordinal encoding (for clearly monotonic ordinal features). All the encoders were fitted within the training fold and transformed into a validation fold and test set to prevent information leakage. For text channels, lightweight cleaning and normalization were performed on dynamic electrocardiogram reports, echocardiogram reports, and surgical records, including special symbol unification, unit standardization, date and identifier de-identification, and medical abbreviation expansion; subsequently, fragment-based sentence segmentation and keyword localization were used to enhance key point density.

To implement LLM embedding + structured CNN late fusion, we construct four parallel text encoders, each fine-tuned from a pre-trained LLM (LLaMA, Phi-2, Mistral, MedGemma). Fine-tuning combines continued pre-training and instruction alignment: first, domain-specific continued pre-training is performed on de-identified dynamic electrocardiogram reports, echocardiogram reports, and surgical records from our institution to improve clinical terminology coverage and syntactic robustness; subsequently, supervised contrastive learning with a classification auxiliary objective is used to moderately update the LLMs. To balance computational power and portability, LoRA/QLoRA is used for low-rank adaptation, freezing most of the lower-layer weights and opening up partial rank parameters in the mid-to-high-layer attention blocks and word embeddings. Text representations are uniformly taken from the penultimate layer's [CLS]-equivalent pooled vector and a token-attention-based weighted average, concatenated to form a 1024-dimensional embedding, and then linearly projected to 256 dimensions to match the representation space of the structured branch. For fair comparison, the four LLMs independently train their respective text encoders and downstream fusion classification heads, while the remaining training and evaluation procedures remain consistent, resulting in four comparable multimodal models.

The structured branch uses ResNet1D as a 1D CNN backbone to learn local interactions and hierarchical features from the 30-dimensional features.Specifically, the structured vectors are stacked in a fixed order to form a "feature sequence" of length 30, which is fed into a network containing three convolutional blocks: Conv1d (channels=32, kernel size=3, stride=1) +

BatchNorm + GELU + MaxPool, followed by a cascade of Conv1d(64, 3) and Conv1d(128, 3). The pooling stride of each layer is controlled to cover different receptive fields and extract cross-feature interactions. The convolutional output is then subjected to global average pooling to obtain a 256-dimensional structured embedding, which is enhanced with dropout and layer normalization to improve generalization. Considering recurrence is a relatively small minority class, a conditional tabular Generative Adversarial Network (GAN) is established within the training fold for the structured branch to perform data augmentation and balance the class distribution. We adopt a conditional WGAN-GP variant adapted to tabular data, where the generator is conditioned on class labels and encoded categorical features to generate synthetic positive samples consistent with the real joint distribution. The discriminator is trained with a Lipschitz constraint and gradient penalty to improve stability. To prevent the augmented data from introducing distribution drift and unreasonable feature combinations, we apply a triple screening process after generation: first, density filtering based on Mahalanobis distance to remove low-density outliers; second, hard constraints based on clinical rules (physiological relationships between indicators and consistency of scoring calculations); and third, an envelope screening of the positive class manifold using a one-class SVM fitted only within the training fold. By performing fold-wise augmentation within the training set, we achieve stratified sampling alignment, while maintaining the natural distribution of the validation and test sets to avoid evaluation bias.

The fusion stage follows a late fusion strategy with attention weighting. In each model instance, the 256-dimensional embedding from the structured branch is concatenated with the corresponding 256-dimensional text embedding from the LLM, forming a 512-dimensional joint representation. This representation is then fed into a multi-head scaled dot-product attention module for learnable cross-modal weighting, with the number of heads set to 4 and the key/value dimension set to 64. A gating mechanism is incorporated to suppress noisy text segments or weakly relevant structured components. The attention output is mapped to the final binary classification logit via a two-layer feedforward network (hidden dimension 256, activation GELU, dropout=0.2), and the cross-entropy loss with label-balanced weights is used as the loss function. Optimization is performed using AdamW (learning rate 2e-4, weight decay 0.01), with cosine annealing and warmup. Training employs stratified k-fold cross-validation (k=5), with patient-level splitting to prevent sample leakage. Within each fold, a validation set is used for early stopping and hyperparameter selection. Evaluation metrics include accuracy, F1-score, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC), and 95% confidence intervals are reported. To ensure comparability across the four LLMs, all non-text side components, optimizer settings, training epochs, and early stopping criteria are kept consistent, with only the encoder being replaced and fine-tuned individually on the text side. During inference, deterministic forwarding with a temperature of 0 is used to obtain stable embeddings, and the maximum text length and truncation strategy are fixed to avoid bias caused by differences in context length between models. The network architecture of the model is depicted in **Figure 5B**.

For the explainability analysis, we calculated the SHAP values in the log-odds domain of the fused model output to achieve global and individual explanations. For the structured branch, DeepSHAP was used to approximate the marginal contribution of the convolutional pathway to

the output, reporting the global importance and interaction effects of each original clinical feature. For the text branch, we combined the attention-based token importance with SHAP's text-masking estimation to locate the descriptive words driving the prediction. Based on the probability output of the optimal model, we determined the optimal threshold using Youden's J statistic to classify patients into high- and low-risk groups, followed by a survival analysis to compare the outcome differences between the two groups.

Sample size calculation

To evaluate the impact of sample size on model stability, we performed additional analyses on the optimal model. In machine learning, Faber and Fonseca demonstrated that increasing the sample size beyond a certain range may not significantly improve outcomes [19]. As this was a retrospective study, the number of patients extracted from each center's electronic hospital information system within a predefined time frame was fixed. Therefore, with reference to previous literature and considering the sample size of this study's training cohort [20], we independently evaluated the change in the AUC of the optimal model between the training and test sets across sample sizes ranging from 800 to 1500.

Statistical Analysis

Continuous variables with a skewed distribution were expressed as the median and interquartile range (IQR). Normally distributed continuous variables were expressed as the mean ± standard deviation (SD). Categorical variables were summarized as frequencies (n) and percentages (%). DCA was performed to evaluate the clinical value of the model by quantifying the net benefit at different threshold probabilities. K-M curves were plotted for patients in the high- and low-risk groups, and the log-rank test was used to determine whether there was a statistically significant difference in the progression-free survival curves between the two groups. All statistical tests were two-sided, and significance was defined as $p < 0.05$.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available as they contain confidential patient information but are available from the corresponding author upon reasonable request.

Abbreviations

ADASYN Adaptive Synthetic Sampling
AF Atrial Fibrillation
ALP Alkaline Phosphatase
ALT Alanine Aminotransferase
AST Aspartate Aminotransferase
CA Catheter ablation
CNN Convolutional Neural Network
eGFR estimated Glomerular Filtration Rate
FPG Fasting Plasma Glucose
GAN Generative Adversarial Network

HbA1C Hemoglobin A1c
HDL High-Density Lipoprotein
LAD Left Atrial Diameter
LDL Low-Density Lipoprotein
LLMs Large language models
LVEF Left Ventricular Ejection Fraction
WGAN-GP Wasserstein GAN with gradient penalty
MICE multiple imputation chained equations
SHAP SHapley Additive exPlanations
SMOTE synthetic minority oversampling technique
SVM Support Vector Machine
TC Total Cholesterol
TG Triglycerides

References

1. Writing Committee Members et al. 2023 ACC/AHA/ACCP/HRS Guideline for the Diagnosis and Management of Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. J. Am. Coll. Cardiol. 83, 109–279 (2024).
2. Zink, M. D. et al. Predictors of recurrence of atrial fibrillation within the first 3 months after ablation. Europace 22, 1337–1344 (2020).
3. Dretzke, J. et al. Predicting recurrent atrial fibrillation after catheter ablation: a systematic review of prognostic models. Europace 22, 748–760 (2020).
4. Jia, S. et al. Association between triglyceride-glucose index trajectories and radiofrequency ablation outcomes in patients with stage 3D atrial fibrillation. Cardiovasc. Diabetol. 23, 121 (2024).
5. Black-Maier, E. et al. Predicting atrial fibrillation recurrence after ablation in patients with heart failure: Validity of the APPLE and CAAP-AF risk scoring systems. Pacing Clin. Electrophysiol. 42, 1440–1447 (2019).
6. Shool, S. et al. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med. Inform. Decis. Mak. 25, 117 (2025).
7. Google. MedGemma. GitHub repository https://github.com/Google-Health/medgemma (2025).
8. Microsoft. Phi-2: The surprising power of small language models. Microsoft Research Blog https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/ (2023).
9. Touvron, H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288 (2023).
10. Jiang, A. Q. et al. Mistral 7B. arXiv preprint arXiv:2310.06825 (2023).
11. Bedi, S. et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA 333, 319–328 (2025).
12. Parameswaran, R., Al-Kaisey, A. M. & Kalman, J. M. Catheter ablation for atrial fibrillation: current indications and evolving technologies. Nat. Rev. Cardiol. 18, 210–225 (2021).

13. Brundel, B. J. J. M. et al. Atrial fibrillation. Nat. Rev. Dis. Primers 8, 21 (2022).
14. Hu, Y. F., Chen, Y. J., Lin, Y. J. & Chen, S. A. Inflammation and the pathogenesis of atrial fibrillation. Nat. Rev. Cardiol. 12, 230–243 (2015).
15. Krasteva, V. et al. Detection of Atrial Fibrillation in Holter ECG Recordings by ECHOView Images: A Deep Transfer Learning Study. Diagnostics 15, 865 (2025).
16. Zhang, P. et al. Automatic screening of patients with atrial fibrillation from 24-h Holter recording using deep learning. Eur. Heart J. Digit. Health 4, 216–224 (2023).
17. Jia, S. et al. A Simple Logistic Regression Model for Predicting the Likelihood of Recurrence of Atrial Fibrillation Within 1 Year After Initial Radio-Frequency Catheter Ablation Therapy. Front. Cardiovasc. Med. 8, 819341 (2022).
18. Sohns, C. & Marrouche, N. F. Atrial fibrillation and cardiac fibrosis. Eur. Heart J. 41, 1123–1131 (2020).
19. Faber, J. & Fonseca, L. M. How sample size influences research outcomes. Dental Press J. Orthod. 19, 27–29 (2014).
20. Rajput, D., Wang, W. J. & Chen, C. C. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics 24, 48 (2023).

Figure 1. Comparison of Metrics Among Different Models

Figure 1A. Radar chart comparing different LLM-based text extraction modules in the dual-branch deep learning network architecture on the training set.
Figure 1B. Radar chart comparing different LLM-based text extraction modules in the dual-branch deep learning network architecture on the validation set.
Figure 1C. Radar chart comparing different LLM-based text extraction modules in the dual-branch deep learning network architecture on the test set.
Figure 1D. Receiver operating characteristic (ROC) curves of different prediction models on the training set.
Figure 1E. Receiver operating characteristic (ROC) curves of different prediction models on the validation set.
Figure 1F. Receiver operating characteristic (ROC) curves of different prediction models on the test set.
Figure 1G. Decision curve analysis (DCA) of different prediction models on the training set.
Figure 1H. Decision curve analysis (DCA) of different prediction models on the validation set.
Figure 1I. Decision curve analysis (DCA) of different prediction models on the test set.
Figure 1J. Bubble chart of DeLong's test for different prediction models on the training set.
Figure 1K. Bubble chart of DeLong's test for different prediction models on the validation set.
Figure 1L. Bubble chart of DeLong's test for different prediction models on the test set.

Figure 2. Kaplan-Meier Survival Curves for the Optimal Model

Figure 2A. Kaplan-Meier survival curves for the training set.
Figure 2B. Kaplan-Meier survival curves for the validation set.
Figure 2C. Kaplan-Meier survival curves for the test set.

Figure 3. Model Interpretability

Figure 3A. SHAP variable importance plot for the structured data channel.
Figure 3B. SHAP token-level importance plot from the text channel.
Figure 3C. Word cloud visualization of key features extracted from the text channel.

Figure 4. Forest Plot of Sensitivity Analysis

Forest plot displaying the results of sensitivity analysis across different subgroups and conditions.

Figure 5. Flow Chart of this Study

Figure 5A. Study enrollment flowchart illustrating participant selection and exclusion criteria.
Figure 5B. Schematic diagram of the dual-branch deep learning network architecture showing the integration of structured data and text processing channels.

Author Contributions

Conceptualization：J.S., Y.Y, X.S.
Methodology：J.S., L.C.
Software：J.S., L.C.
Validation：Z.Q.,Y.Y., Z.H., Z.J.
Formal analysis: J.S.
Investigation：M.X.. Y.Q., Z.J.
Resources：Y.Y., G.Y., Y.X., S.P., L.J., X.Q.
Data curation：J.S, Y.Y., G.Y., Y.X. ,S.P.
Writing—original draft：J.S.
Writing—review and editing： J.S., X.S.
Visualization：L.W, F.C., Y.Y.

Supervision：X.S.,G.W.
Project administration：X.S.,G.W.
Funding acquisition：J.S., X.S., Y.Y.

Funding

This study was supported by Key Projects of the Zhejiang Provincial Natural Science Foundation (No.LZ25H020001), Joint TCM Science & Technology Projects of National Demonstration Zones for Comprehensive TCM Reformt (No. GZY-KJS-ZJ-2025-022), National Natural Science Foundation of China (No.81971688), Scientific Research Fund of Zhejiang Provincial Education Department (No.Y202457057) and Taizhou Science and Technology Project (No.21ywa02).

Consent for publication

Not applicable.

Acknowledgments

We are grateful to PixelmedAI platform (https://github.com/410312774/PixelMedAI) for providing technical support for deep learning network in this study.

Conflict of Interest

Authors declare that they have no competing interests.

Table 1. Subjects Baseline

Variables	**Total (n = 2508)**	**ZJU4th (n=714)**	ZJTZH **(n=1002)**	YNH **(n=442)**	JHPH **(n=221)**	NBH **(n=129)**	**Statistic**	**P**
Age (years), M (Q_1, Q_3)	65.00 (58.00, 71.00)	68.00 (61.00,72.00)	64.00 (57.00,70.00)	64.00 (56.00,71.00)	66.00 (59.00,72.00)	67.00 (58.00,74.00)	χ^2=56.42#	<.01
BMI (kg/m2), M (Q_1, Q_3)	24.62 (23.12, 26.25)	24.72 (23.50,26.02)	24.64 (22.59,26.77)	24.55 (22.50,26.56)	24.47 (23.92,25.05)	25.02 (23.23,27.02)	χ^2=6.50#	0.16
Gender, n(%)							χ^2=10.98	0.03
Male	1572 (62.68)	445 (62.32)	634 (63.27)	267 (60.41)	156 (70.59)	70 (54.26)		
Female	936 (37.32)	269 (37.68)	368 (36.73)	175 (39.59)	65 (29.41)	59 (45.74)		
Hypertension, n(%)							χ^2=34.95	<.01

None	1148 (45.77)	284 (39.78)	519 (51.80)	193 (43.67)	109 (49.32)	43 (33.33)		
Yes	1360 (54.23)	430 (60.22)	483 (48.20)	249 (56.33)	112 (50.68)	86 (66.67)		
Coronary Artery Disease, n(%)							χ^2=35.64	<.01
None	2037 (81.22)	578 (80.95)	848 (84.63)	317 (71.72)	185 (83.71)	109 (84.50)		
Yes	471 (18.78)	136 (19.05)	154 (15.37)	125 (28.28)	36 (16.29)	20 (15.50)		
Diabetes, n(%)							χ^2=25.76	<.01
None	2086 (83.17)	588 (82.35)	872 (87.03)	358 (81.00)	164 (74.21)	104 (80.62)		
Yes	422 (16.83)	126 (17.65)	130 (12.97)	84 (19.00)	57 (25.79)	25 (19.38)		
AF type, n(%)							χ^2=113.41	<.01
Paroxysmal	1542 (61.48)	455 (63.73)	535 (53.39)	267 (60.41)	202 (91.40)	83 (64.34)		
Persistent	966 (38.52)	259 (36.27)	467 (46.61)	175 (39.59)	19 (8.60)	46 (35.66)		
Status, n(%)							χ^2=30.55	<.01
None	1942 (77.43)	526 (73.67)	827 (82.53)	315 (71.27)	172 (77.83)	102 (79.07)		
Recurrence	566 (22.57)	188 (26.33)	175 (17.47)	127 (28.73)	49 (22.17)	27 (20.93)		
Systolic Blood Pressure (mmHg), M (Q_1, Q_3)	127.80 (117.00, 139.00)	128.00 (116.00,141.00)	127.95 (115.00,140.00)	120.00 (114.00,138.00)	128.00 (124.00,132.60)	131.00 (118.00,145.00)	χ^2=20.43#	<.01
Diastolic Blood Pressure (mmHg), M (Q_1, Q_3)	79.40 (71.00, 86.00)	78.00 (70.00,85.00)	79.00 (72.00,88.00)	80.00 (70.00,84.00)	78.60 (76.70,81.20)	82.00 (73.00,89.00)	χ^2=18.71#	<.01
AF Duration (months), M (Q_1, Q_3)	12.00 (2.00, 48.00)	3.00 (0.50,24.00)	12.00 (3.00,48.00)	24.02 (5.79,72.00)	32.15 (20.16,48.15)	5.00 (1.00,24.00)	χ^2=246.68#	<.01
CHA2DS2-VASc, Mean ± SD	2.32 ± 1.51	2.27 ± 1.42	2.38 ± 1.60	2.19 ± 1.60	2.04 ± 0.66	3.02 ± 1.67	F=10.34	<.01
HAS-BLED, Mean ± SD	0.77 ± 0.81	1.18 ± 0.94	0.61 ± 0.72	0.55 ± 0.62	0.59 ± 0.59	0.77 ± 0.82	F=74.01	<.01
LAD (mm), M (Q_1, Q_3)	39.00 (35.00, 43.00)	37.00 (33.00,41.30)	41.00 (36.00,45.00)	39.00 (35.00,43.00)	38.00 (33.00,40.00)	41.00 (37.00,44.00)	χ^2=154.98#	<.01

LVEF(%), M (Q_1, Q_3)	62.28 (58.00, 67.00)	64.00 (60.00,68.30)	60.00 (54.00,65.00)	64.00 (59.00,68.00)	62.25 (60.16,63.25)	65.00 (62.00,68.00)	χ^2=183.84#	<.001
Survival Time (months), M (Q_1, Q_3)	31.00 (17.00, 48.00)	25.00 (14.00,41.00)	32.00 (19.00,53.00)	43.00 (17.00,69.00)	34.40 (25.30,39.90)	25.00 (17.00,33.00)	χ^2=125.29#	<.001
HbA1C, M (Q_1, Q_3)	6.10 (5.80, 6.44)	6.06 (5.70,6.46)	6.10 (5.80,6.40)	6.20 (5.78,6.60)	6.11 (6.02,6.28)	6.10 (5.80,6.40)	χ^2=12.40#	.015
FPG (mmol/L), M (Q_1, Q_3)	5.14 (4.58, 5.87)	4.91 (4.44,5.48)	5.55 (5.03,6.53)	4.33 (2.49,5.21)	5.36 (5.04,5.72)	4.69 (4.26,5.15)	χ^2=496.28#	<.001
TC (mmol/L), M (Q_1, Q_3)	4.08 (3.49, 4.77)	3.87 (3.25,4.54)	4.36 (3.65,5.06)	4.11 (3.41,4.84)	3.99 (3.73,4.15)	4.04 (3.25,4.61)	χ^2=95.25#	<.001
TG (mmol/L), M (Q_1, Q_3)	1.36 (0.99, 1.90)	1.24 (0.93,1.73)	1.32 (0.94,1.87)	1.64 (1.08,4.46)	1.43 (1.29,1.57)	1.30 (0.97,1.92)	χ^2=93.16#	<.001
HDL (mmol/L), M (Q_1, Q_3)	1.16 (0.99, 1.35)	1.08 (0.91,1.27)	1.22 (1.07,1.45)	1.11 (0.93,1.27)	1.26 (1.10,1.55)	1.11 (0.94,1.25)	χ^2=187.09#	<.001
LDL (mmol/L), M (Q_1, Q_3)	2.24 (1.77, 2.76)	2.12 (1.60,2.65)	2.34 (1.82,2.86)	2.40 (1.80,3.06)	2.16 (1.96,2.31)	2.32 (1.66,2.81)	χ^2=59.53#	<.001
Albumin(g/L), M (Q_1, Q_3)	39.90 (37.50, 43.18)	38.60 (36.60,40.80)	42.10 (38.10,45.00)	40.90 (37.70,43.80)	39.22 (38.41,39.76)	39.70 (37.50,41.50)	χ^2=237.81#	<.001
ALT (U/L), M (Q_1, Q_3)	22.00 (15.00, 30.07)	19.00 (14.00,28.00)	23.00 (16.00,35.00)	23.00 (16.00,32.00)	23.10 (21.50,28.10)	17.00 (14.00,25.00)	χ^2=87.62#	<.001
AST (U/L), M (Q_1, Q_3)	23.00 (19.00,	23.00 (19.00,28.	24.00 (20.00,31.	22.00 (18.00,28.	24.70 (22.70,26.	19.00 (16.00,23.	χ^2=124.3	<.

	28.25)	00)	00)	00)	70)	00)	5#	0 0 1
ALP (U/L), M (Q_1, Q_3)	76.00 (64.00, 88.00)	73.45 (61.00,87.00)	78.00 (65.00,92.00)	74.00 (61.00,90.00)	76.40 (69.80,80.50)	68.00 (57.00,83.00)	χ^2=35.24#	<.001
Creatine (umol/L), M (Q_1, Q_3)	74.89 (65.00, 85.00)	73.00 (62.00,86.00)	76.00 (66.00,87.00)	75.00 (65.25,88.00)	75.50 (73.40,79.30)	67.00 (56.00,78.00)	χ^2=46.46#	<.01
eGFR (ml/min/1.73m²), M (Q_1, Q_3)	87.00 (75.00, 96.00)	90.00 (78.00,103.00)	84.00 (71.00,94.00)	86.25 (72.90,97.10)	88.70 (84.10,93.10)	86.87 (70.03,95.67)	χ^2=88.09#	<.01
APPLE, Mean ± SD	1.89 ± 1.16	1.75 ± 1.14	2.13 ± 1.18	1.74 ± 1.19	1.59 ± 0.69	1.88 ± 1.25	F=19.16	<.01
CAAP-AF, Mean ± SD	3.98 ± 2.04	3.90 ± 1.93	4.23 ± 2.09	3.94 ± 2.21	2.99 ± 1.39	4.29 ± 2.02	F=18.36	<.01
I/III Anti-arrythmia Drug Use, n(%)							χ^2=65.07	<.01
None	505 (20.14)	130 (18.21)	276 (27.54)	54 (12.22)	26 (11.76)	19 (14.73)		
Yes	2003 (79.86)	584 (81.79)	726 (72.46)	388 (87.78)	195 (88.24)	110 (85.27)		
II Anti-arrythmia Drug Use, n(%)							χ^2=22.66	<.01
None	2001 (79.78)	538 (75.35)	793 (79.14)	371 (83.94)	194 (87.78)	105 (81.40)		
Yes	507 (20.22)	176 (24.65)	209 (20.86)	71 (16.06)	27 (12.22)	24 (18.60)		

#: Kruskal-waills test, χ^2: Chi-square test, F: ANOVA

M: Median, Q_1: 1st Quartile, Q_3: 3st Quartile , SD: standard deviation

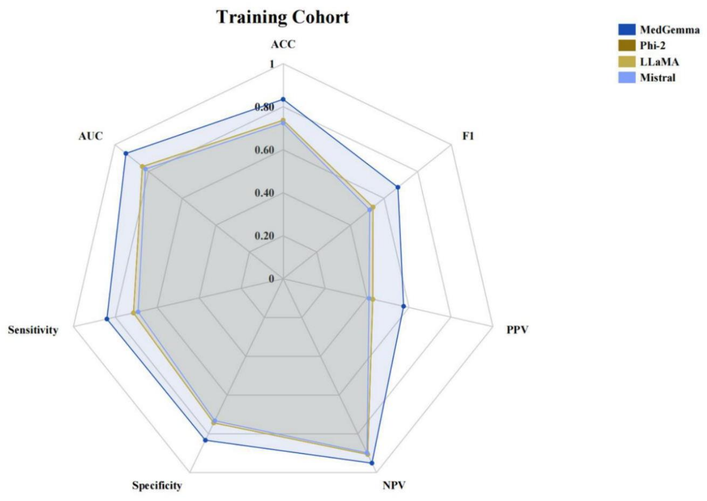
Table 2. Model Performance Comparison

Model	Dataset	ACC	AUC	Sensitivity	Specificity	NPV	PPV	F1
AF type								

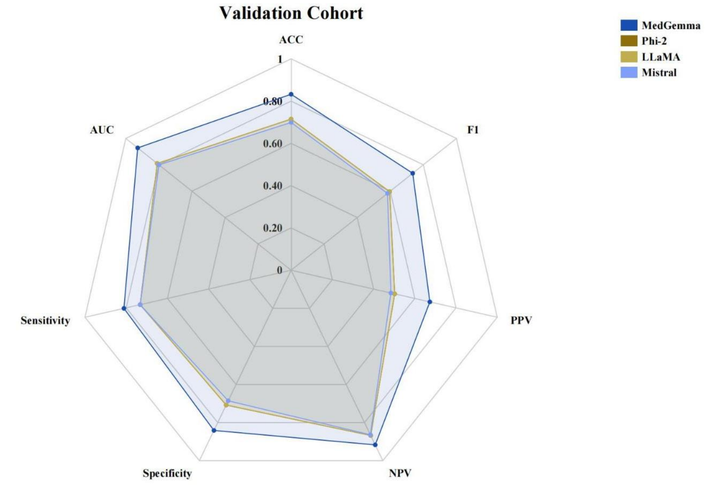
Model	Dataset	ACC	AUC	Sensitivity	Specificity	NPV	PPV	F1
	Train	0.614	0.604	0.587	0.621	0.848	0.293	0.391
	Val	0.656	0.637	0.591	0.683	0.805	0.429	0.497
	Test	0.683	0.507	0.197	0.818	0.786	0.231	0.213
APPLE								
	Train	0.473	0.607	0.758	0.397	0.859	0.252	0.378
	Val	0.624	0.739	0.795	0.556	0.871	0.419	0.549
	Test	0.700	0.473	0.184	0.843	0.788	0.246	0.211
CAAP-AF								
	Train	0.609	0.618	0.581	0.616	0.846	0.289	0.386
	Val	0.733	0.798	0.669	0.759	0.851	0.528	0.590
	Test	0.723	0.516	0.171	0.876	0.792	0.277	0.211
CHA2DS2-VASc								
	Train	0.415	0.519	0.697	0.339	0.807	0.221	0.335
	Val	0.713	0.772	0.701	0.717	0.856	0.500	0.584
	Test	0.329	0.504	0.908	0.168	0.868	0.232	0.370
LAD								
	Train	0.650	0.616	0.529	0.683	0.844	0.309	0.390
	Val	0.724	0.729	0.575	0.784	0.821	0.518	0.545
	Test	0.611	0.613	0.684	0.591	0.871	0.317	0.433
LLaMA-Fusion								

Model	Dataset	ACC	AUC	Sensitivity	Specificity	NPV	PPV	F1
	Train	0.737	0.837	0.713	0.744	0.906	0.427	0.535
	Val	0.715	0.810	0.732	0.708	0.868	0.503	0.596
	Test	0.723	0.828	0.763	0.712	0.915	0.423	0.545
Mistral-Fusion								
	Train	0.724	0.819	0.691	0.732	0.898	0.409	0.514
	Val	0.699	0.800	0.732	0.686	0.864	0.484	0.583
	Test	0.674	0.777	0.711	0.664	0.892	0.370	0.486
Phi-2 Fustion								
	Train	0.737	0.837	0.713	0.744	0.906	0.427	0.535
	Val	0.715	0.810	0.732	0.708	0.868	0.503	0.596
	Test	0.697	0.795	0.737	0.686	0.904	0.394	0.514
MedGemma Fusion								
	Train	0.834	0.935	0.840	0.833	0.951	0.574	0.682
	Val	0.833	0.928	0.811	0.841	0.917	0.673	0.736
	Test	0.809	0.912	0.816	0.807	0.940	0.539	0.649

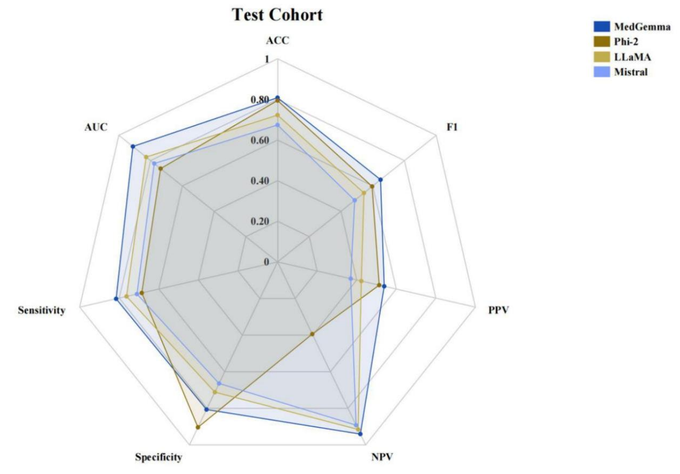
A.



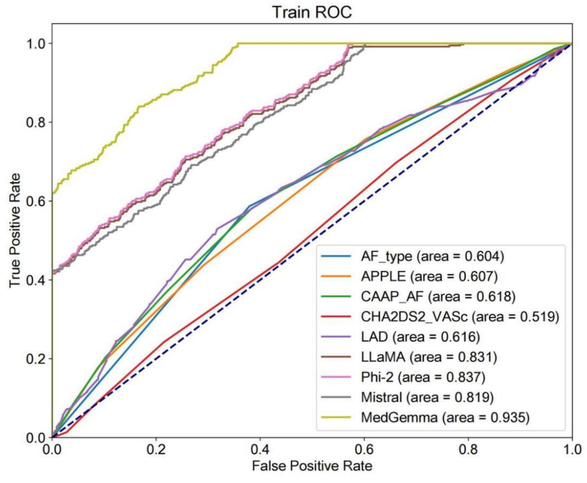
B.



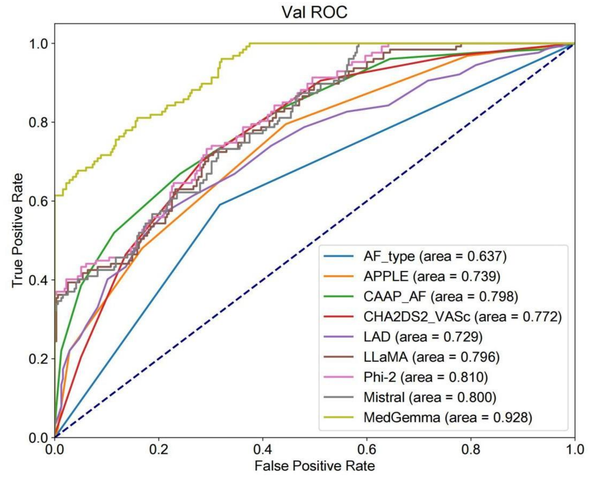
C.



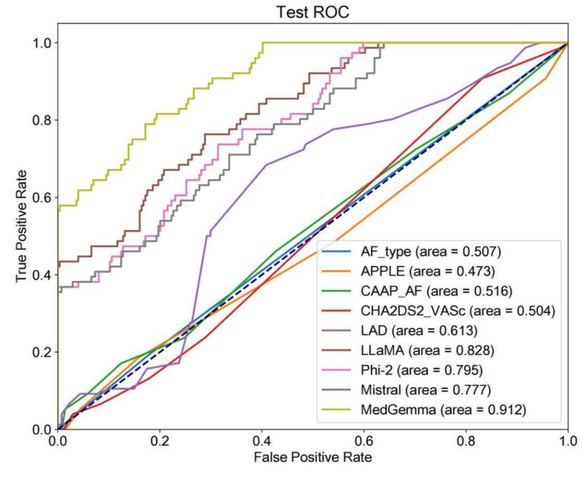
D.



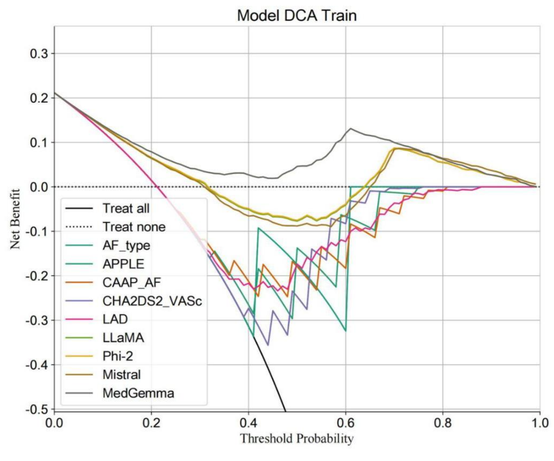
E.



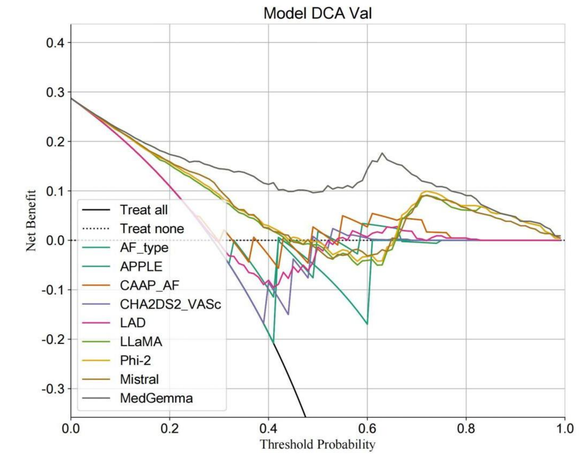
F.



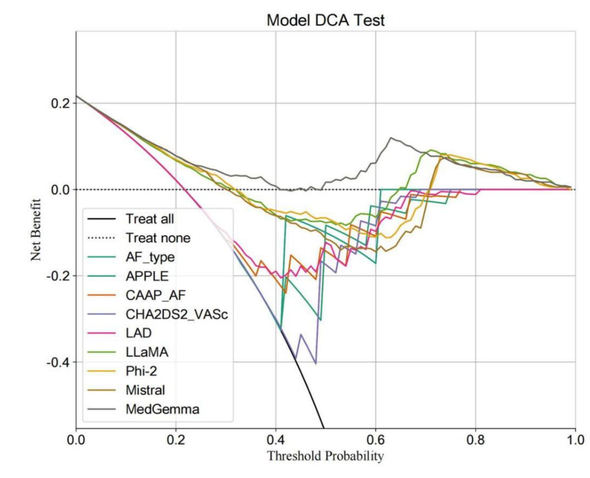
G.



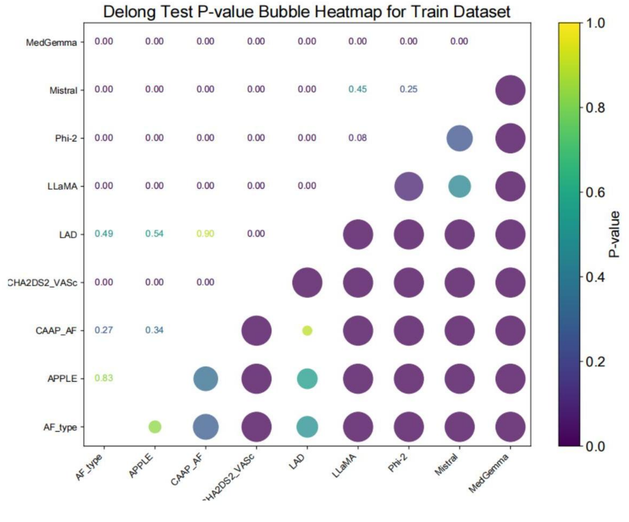
H.



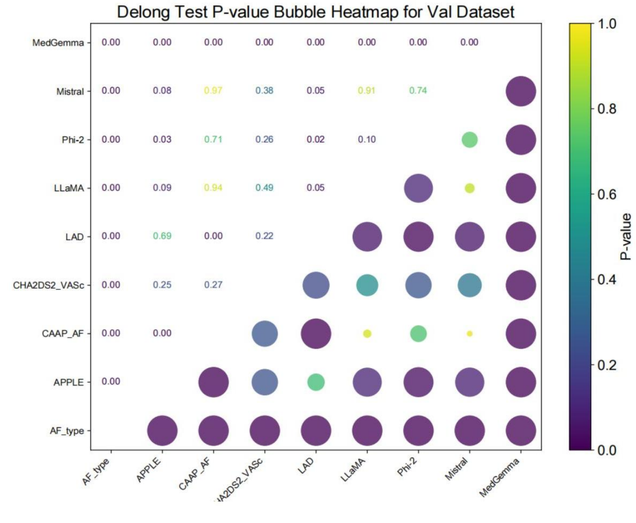
I.



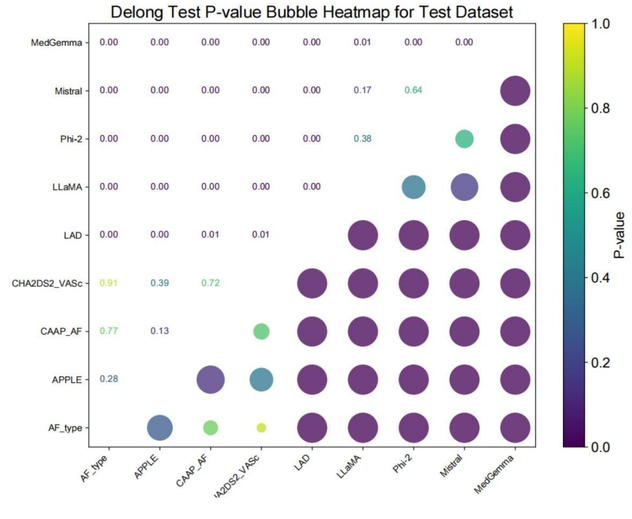
J.



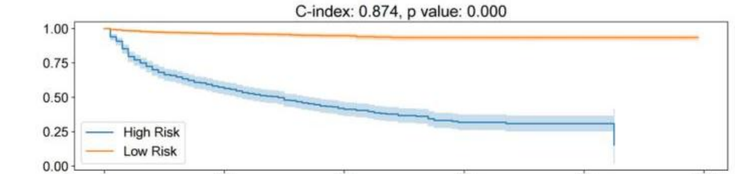
K.



L.

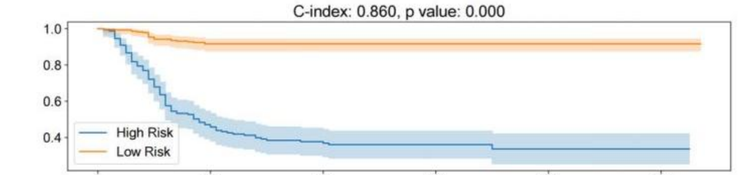


A.



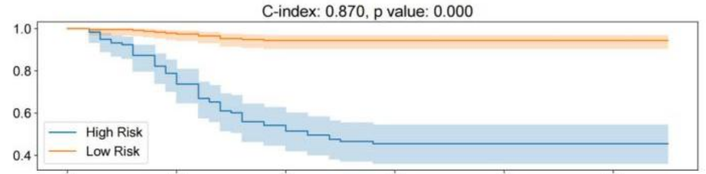
High Risk						
At risk	531	253	110	43	5	0
Censored	0	50	137	186	223	227
Events	0	228	284	302	303	304
Low Risk						
At risk	1185	883	455	179	25	0
Censored	0	258	677	947	1101	1126
Events	0	44	53	59	59	59

B.



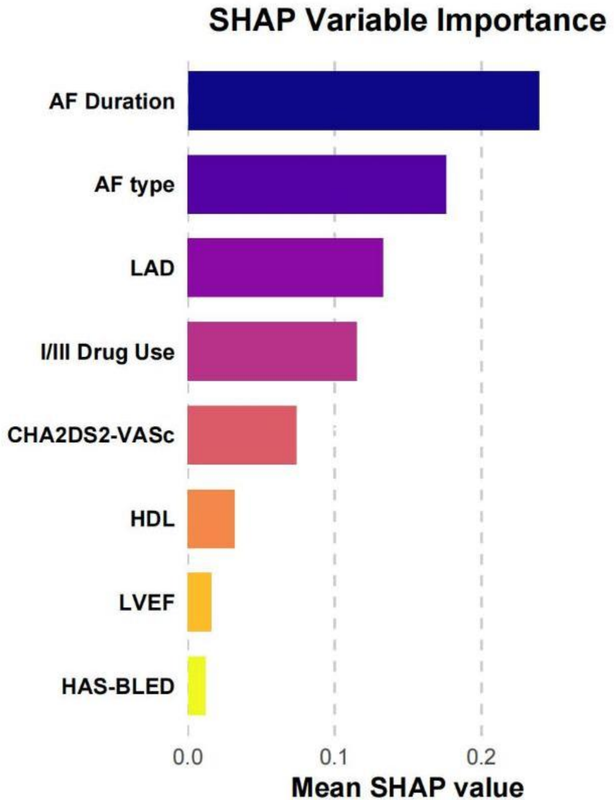
High Risk						
At risk	165	72	46	23	9	1
Censored	0	4	17	39	52	60
Events	0	89	102	103	104	104
Low Risk						
At risk	277	235	190	121	63	13
Censored	0	19	64	133	191	241
Events	0	23	23	23	23	23

C.

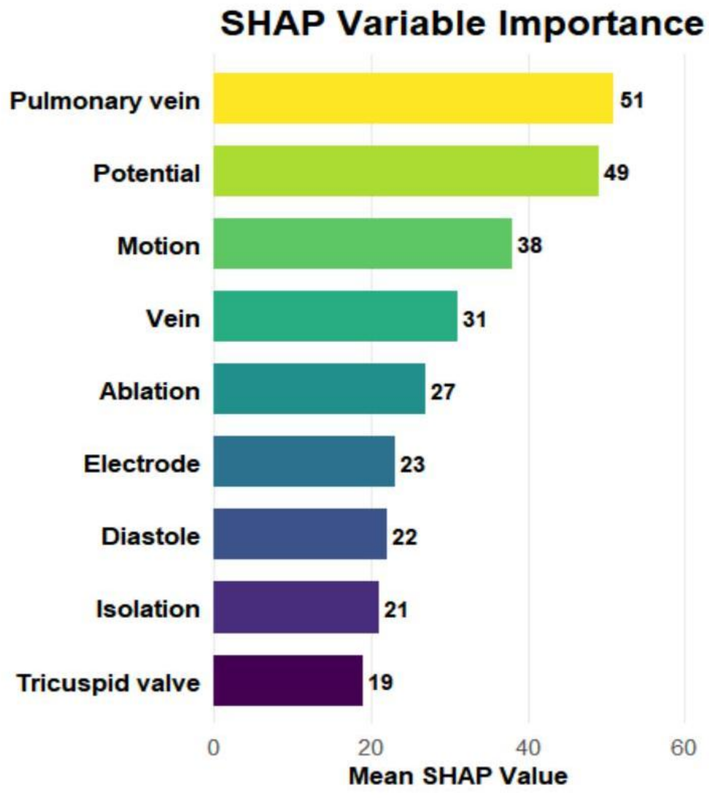


High Risk						
At risk	118	87	54	42	20	2
Censored	0	0	7	13	35	53
Events	0	31	57	63	63	63
Low Risk						
At risk	232	226	196	150	55	3
Censored	0	0	23	69	164	216
Events	0	6	13	13	13	13

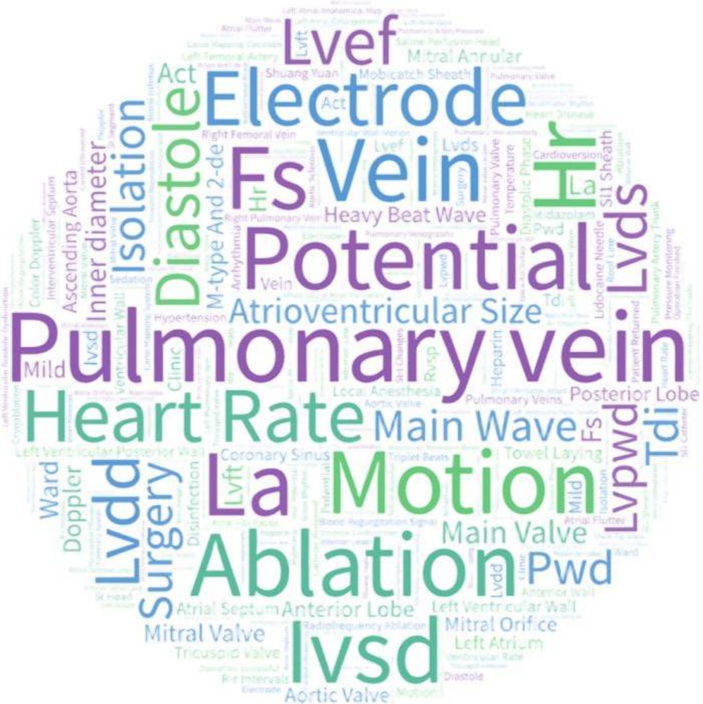
A.

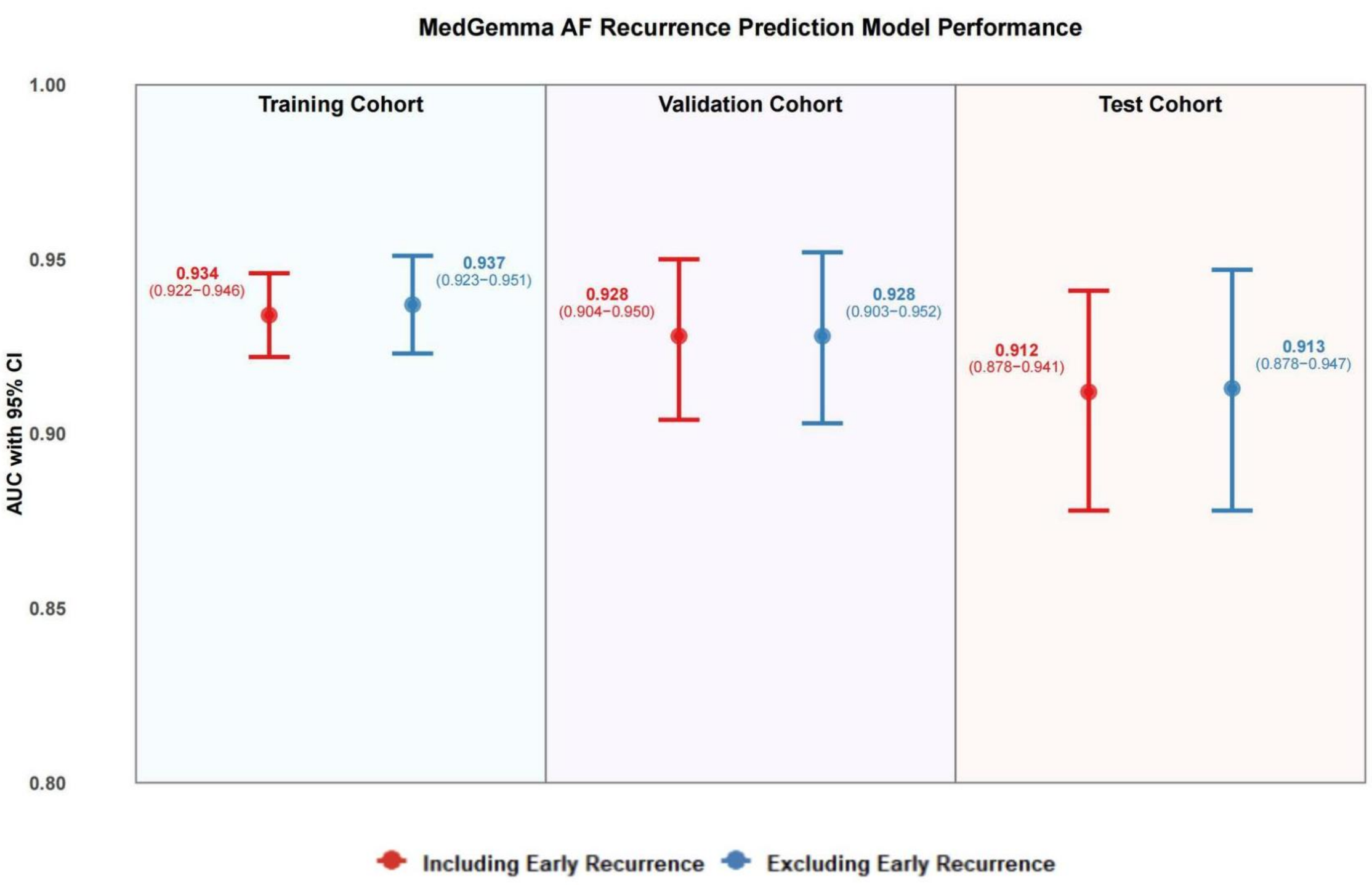

SHAP Variable Importance
AF Duration
AF type
LAD
I/III Drug Use
CHA2DS2-VASc
HDL
LVEF
HAS-BLED
0.0
0.1
0.2
Mean SHAP value

B.

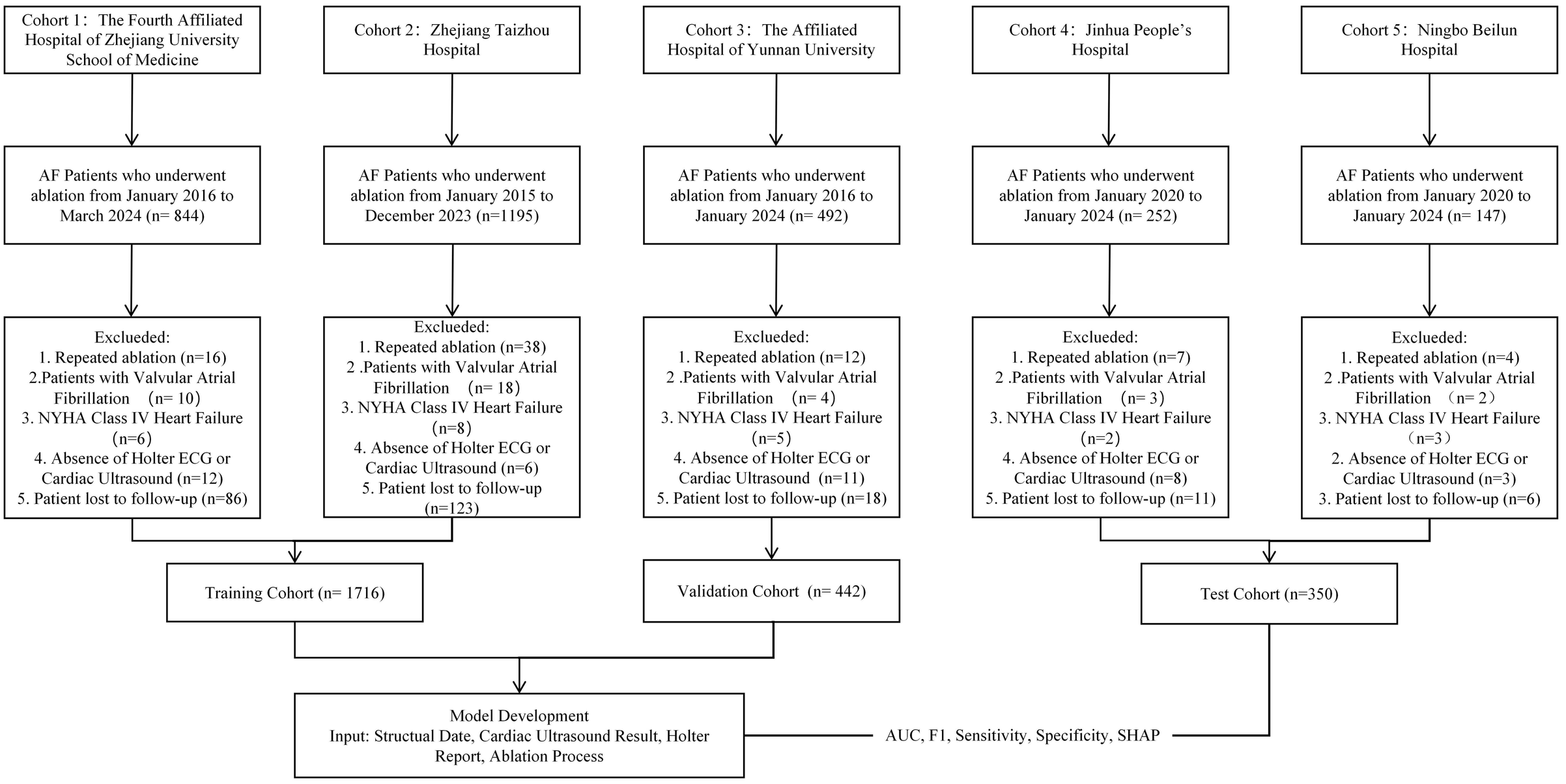

SHAP Variable Importance
Pulmonary vein
Potential
Motion
Vein
Ablation
Electrode
Diastole
Isolation
Tricuspid valve
51
49
38
31
27
23
22
21
19
0
20
40
60
Mean SHAP Value

C.

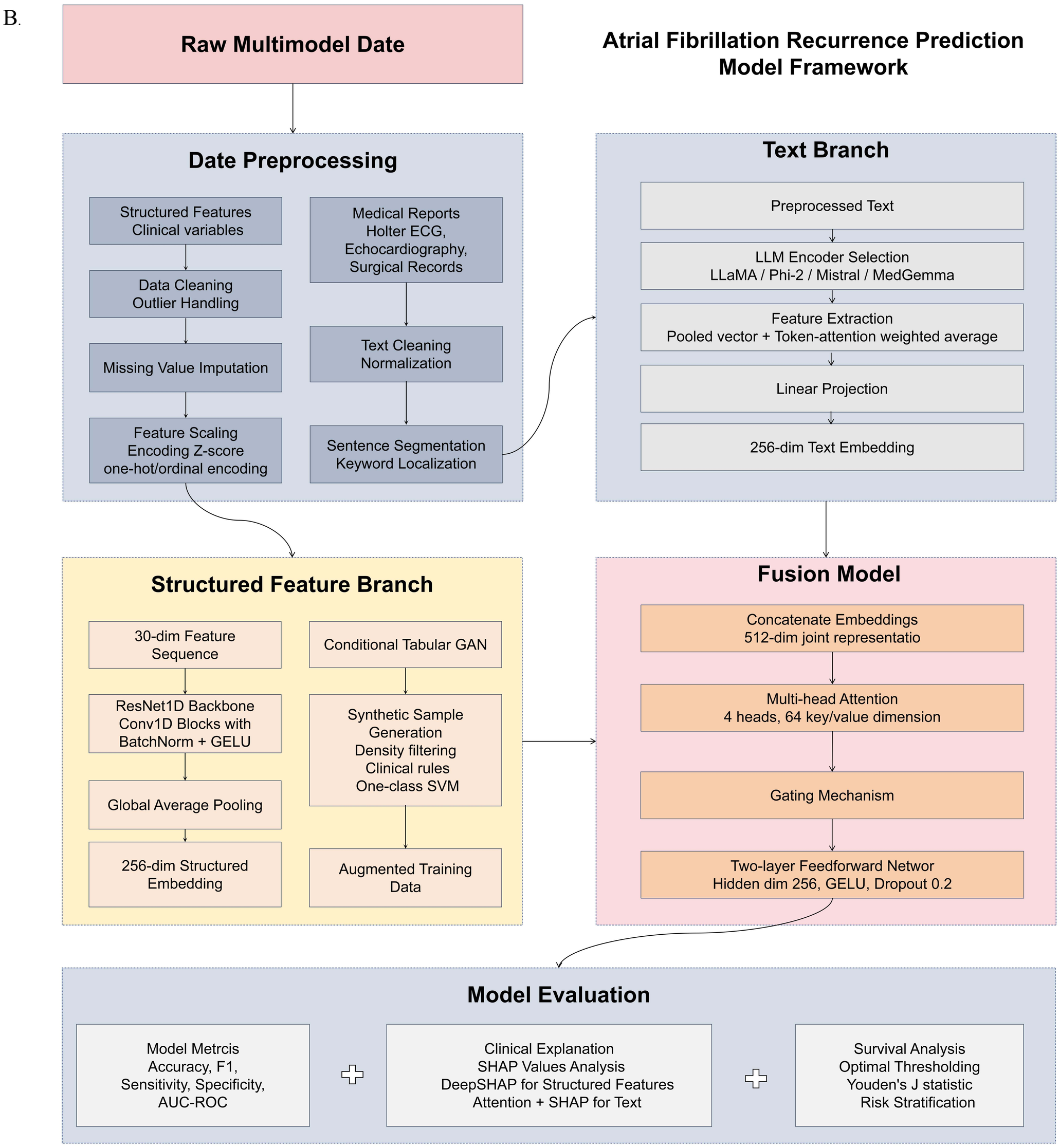

Lvef
Electrode
Fs
Vein
Hr
Potential
Diastole
Isolation
Lvds
Pulmonary vein
Heart Rate
Main Wave
Tdi
Lvpwd
La
Motion
Lvdd
Surgery
Ablation
Pwd
Main Valve
Ivsd
Atrioventricular Size
Heavy Beat Wave
Inner diameter
Ascending Aorta
Posterior Lobe
Anterior Lobe
Mitral Valve
Aortic Valve
Left Atrium
Mitral Orifice


MedGemma AF Recurrence Prediction Model Performance
AUC with 95% CI
1.00
0.95
0.90
0.85
0.80
Training Cohort
Validation Cohort
Test Cohort
0.934
(0.922–0.946)
0.937
(0.923–0.951)
0.928
(0.904–0.950)
0.928
(0.903–0.952)
0.912
(0.878–0.941)
0.913
(0.878–0.947)
Including Early Recurrence
Excluding Early Recurrence

A.


Cohort 1：The Fourth Affiliated Hospital of Zhejiang University School of Medicine
Cohort 2：Zhejiang Taizhou Hospital
Cohort 3：The Affiliated Hospital of Yunnan University
Cohort 4：Jinhua People's Hospital
Cohort 5：Ningbo Beilun Hospital
AF Patients who underwent ablation from January 2016 to March 2024 (n= 844)
AF Patients who underwent ablation from January 2015 to December 2023 (n=1195)
AF Patients who underwent ablation from January 2016 to January 2024 (n= 492)
AF Patients who underwent ablation from January 2020 to January 2024 (n= 252)
AF Patients who underwent ablation from January 2020 to January 2024 (n= 147)
Excluded:
1. Repeated ablation (n=16)
2.Patients with Valvular Atrial Fibrillation （n= 10）
3. NYHA Class IV Heart Failure (n=6)
4. Absence of Holter ECG or Cardiac Ultrasound (n=12)
5. Patient lost to follow-up (n=86)
Excluded:
1. Repeated ablation (n=38)
2 .Patients with Valvular Atrial Fibrillation （n= 18）
3. NYHA Class IV Heart Failure (n=8)
4. Absence of Holter ECG or Cardiac Ultrasound (n=6)
5. Patient lost to follow-up (n=123)
Excluded:
1. Repeated ablation (n=12)
2 .Patients with Valvular Atrial Fibrillation （n= 4）
3. NYHA Class IV Heart Failure (n=5)
4. Absence of Holter ECG or Cardiac Ultrasound （n=11）
5. Patient lost to follow-up (n=18)
Excluded:
1. Repeated ablation (n=7)
2 .Patients with Valvular Atrial Fibrillation （n= 3）
3. NYHA Class IV Heart Failure (n=2)
4. Absence of Holter ECG or Cardiac Ultrasound (n=8)
5. Patient lost to follow-up (n=11)
Excluded:
1. Repeated ablation (n=4)
2 .Patients with Valvular Atrial Fibrillation （n= 2）
3. NYHA Class IV Heart Failure （n=3）
2. Absence of Holter ECG or Cardiac Ultrasound (n=3)
3. Patient lost to follow-up (n=6)
Training Cohort (n= 1716)
Validation Cohort （n= 442）
Test Cohort (n=350)
Model Development
Input: Structual Date, Cardiac Ultrasound Result, Holter Report, Ablation Process
AUC, F1, Sensitivity, Specificity, SHAP

B.


Raw Multimodel Date
Atrial Fibrillation Recurrence Prediction Model Framework
Date Preprocessing
Structured Features Clinical variables
Data Cleaning Outlier Handling
Missing Value Imputation
Feature Scaling Encoding Z-score one-hot/ordinal encoding
Medical Reports Holter ECG, Echocardiography, Surgical Records
Text Cleaning Normalization
Sentence Segmentation Keyword Localization
Text Branch
Preprocessed Text
LLM Encoder Selection LLaMA / Phi-2 / Mistral / MedGemma
Feature Extraction Pooled vector + Token-attention weighted average
Linear Projection
256-dim Text Embedding
Structured Feature Branch
30-dim Feature Sequence
ResNet1D Backbone Conv1D Blocks with BatchNorm + GELU
Global Average Pooling
256-dim Structured Embedding
Conditional Tabular GAN
Synthetic Sample Generation Density filtering Clinical rules One-class SVM
Augmented Training Data
Fusion Model
Concatenate Embeddings 512-dim joint representatio
Multi-head Attention 4 heads, 64 key/value dimension
Gating Mechanism
Two-layer Feedforward Networ Hidden dim 256, GELU, Dropout 0.2
Model Evaluation
Model Metrcis Accuracy, F1, Sensitivity, Specificity, AUC-ROC
Clinical Explanation SHAP Values Analysis DeepSHAP for Structured Features Attention + SHAP for Text
Survival Analysis Optimal Thresholding Youden's J statistic Risk Stratification