

Patient-Level Out-of-Fold-Guided Threshold-Aware Heterogeneous Convolutional Neural Network–XGBoost Probability Fusion for Identification of Atrial Fibrillation from PTB-XL

Poonam Chhabra

*Manav Rachna International Institute of Research and Studies,
Faridabad, Haryana, India.*

poonam.chhabra14@gmail.com

Rosy Madaan

*Manav Rachna International Institute of Research and Studies,
Faridabad, Haryana, India.*

rosymadaan.set@mriu.edu.in

Corresponding Author: Poonam Chhabra

Copyright © 2026 Poonam Chhabra and Rosy Madaan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Atrial fibrillation (AF) is a prevalent cardiac rhythm disorder that requires accurate and reliable detection from electrocardiogram (ECG) recordings. This study proposes a patient-level, threshold-aware heterogeneous CNN–XGBoost probability-fusion framework that integrates complementary waveform-based and engineered ECG representations for AF detection using the PTB-XL database. The study included 21,799 twelve-lead ECG recordings, comprising 1,514 AF and 20,285 non-AF recordings. A total of 91 ECG-derived temporal, morphological, statistical, and frequency-domain features were used by the XGBoost branch, while a convolutional neural network (CNN) model learned representations directly from preprocessed twelve-lead ECG waveforms. Patient-level partitioning was used to prevent patient overlap between development and independent test sets. Within the development set, strict five-fold out-of-fold (OOF) predictions were generated for model comparison, fusion-weight selection, and decision-threshold selection, while the independent test set remained completely untouched until final evaluation. CNN and XGBoost probabilities were combined using development-OOF-selected weights of 0.75 and 0.25, respectively, and the decision threshold was fixed at 0.740 before independent testing. On 4,429 independent test ECG recordings, the proposed fusion achieved 98.56% accuracy, 90.29% precision, 89.14% sensitivity, 99.27% specificity, 89.71% F1-score, 94.20% balanced accuracy, 0.889 Matthews correlation coefficient (MCC), 0.889 Cohen's kappa, 0.9897 ROC-AUC, and 0.9453 PR-AUC, with a Brier score of 0.0134. Compared with the CNN alone, fusion produced a larger observed increase in PR-AUC than in ROC-AUC, while also yielding a lower Brier score. These findings show that integrating complementary waveform-based and engineered ECG representations was associated with improved precision–recall performance under leakage-controlled patient-level evaluation. External validation using independent multicentre and prospective ECG datasets is required to assess generalizability.

Keywords: Atrial fibrillation, Electrocardiography, Ensemble learning, Machine learning, XGBoost.

1. INTRODUCTION

Atrial fibrillation (AF) is a prevalent sustained cardiac rhythm disorder and is linked to major complications including stroke, heart failure, dementia, and increased mortality. Its occurrence rises with advancing age and the presence of cardiovascular risk factors [1]. Electrocardiography (ECG) remains a principal tool for AF detection because it provides direct information about cardiac electrical activity [1, 2]. However, AF can be difficult to identify when rhythm abnormalities are intermittent or when recordings must be assessed at scale. These challenges have motivated the development of automated ECG interpretation through machine-learning and deep-learning techniques [3, 4].

A broad range of computational approaches has been investigated for automated AF and arrhythmia detection, including conventional machine-learning methods, convolutional neural networks, QRS- and beat-centric approaches, and Transformer-based architectures [4–10]. Studies have also explored AF and atrial flutter detection [11], single-lead ECG analysis [12], PTB-XL-based ECG classification [13–15], resource-efficient monitoring [16], and explainable artificial intelligence [17, 18]. Multi-lead architectures have further demonstrated the value of combining information across ECG leads for arrhythmia detection [19]. Collectively, these studies demonstrate the potential of automated ECG analysis, but they also show that performance depends on the representation of ECG information, class distribution, probability calibration, and evaluation design [4, 20].

A limitation of relying on a single representation is that waveform-based deep-learning models and engineered-feature models represent distinct characteristics of ECG signals. A CNN can derive task-specific representations from multilead waveform signals, whereas engineered features can explicitly describe temporal, morphological, statistical, and frequency-domain properties. Integrating these complementary representations may consequently provide information that may not be fully represented by either modelling approach individually. Ensemble learning offers a suitable mechanism for this integration [21, 22]. However, robust evaluation of an ensemble requires separation at the patient level and strict control of model-development information, especially when a dataset contains multiple ECG recordings from individual patients [23].

The choice of operating threshold is another important consideration used to convert predicted probabilities into AF classifications. A default threshold of 0.50 may not provide a suitable operating point for an imbalanced classification setting, and calibration should be evaluated independently of discrimination [20, 24]. In addition, ROC-AUC alone may not provide a complete characterization of performance when the positive class is relatively uncommon; precision–recall analysis, calibration measures, and additional classification metrics can offer complementary information [4, 20].

The principal methodological contribution of this work lies in combining complementary waveform-based and engineered-feature representations within a patient-level, OOF-controlled probability-fusion framework for AF detection. A CNN processes twelve-lead ECG waveforms, while XGBoost models 91 engineered ECG features. Rather than relying on a logistic-regression stacking layer, the two model probabilities are combined through a development-set OOF-selected weighted fusion. The decision threshold is also selected exclusively from development OOF predictions and fixed before evaluation on an independent patient-level test set. This design separates model development and operating-point selection from final evaluation while directly testing whether the heterogeneous representations provide complementary predictive information.

Accordingly, the specific contributions of this study are:

1. Heterogeneous representation fusion: A CNN–XGBoost framework integrates representations learned directly from twelve-lead ECG waveforms with 91 engineered temporal, morphological, statistical, and frequency-domain ECG features.
2. Leakage-controlled patient-level development: ECG recordings are partitioned at the patient level, and strict five-fold OOF prediction is used within the development data so that each OOF prediction is generated without training on recordings from the corresponding held-out patients.
3. Development-only fusion and threshold selection: Fusion weights and the final decision threshold are selected exclusively from development OOF predictions, with the independent patient-level test set remaining untouched until final evaluation.
4. Comprehensive independent evaluation: The final system is evaluated using discrimination, precision–recall performance, classification measures, calibration, bootstrap confidence intervals, paired statistical comparisons, and SHAP analysis of the engineered-feature XGBoost component.

The resulting framework is intended to provide a rigorous evaluation of whether combining heterogeneous ECG representations can improve AF detection beyond the information captured by an individual modelling branch, while maintaining strict separation between development and independent testing.

2. MATERIALS AND METHODS

2.1 Overall Framework

The developed framework integrates two heterogeneous representations of twelve-lead ECG recordings: a waveform-based representation learned by a convolutional neural network (CNN) and an engineered-feature representation modeled by XGBoost. The CNN operates directly on the preprocessed ECG waveform, whereas XGBoost uses 91 engineered ECG features describing temporal, morphological, statistical, and frequency-domain characteristics. The two branches therefore process complementary representations rather than identical inputs.

The experimental workflow consists of patient-level partitioning, fold-specific data processing and model training, five-fold out-of-fold (OOF) prediction within the development set, CNN and XGBoost model training, development-set probability fusion, development-only decision-threshold selection, and final evaluation on an independent patient-level test set. All operations involving model comparison, fusion-weight selection, and threshold selection are restricted to the development data. The independent test set is not used during model development or selection. FIGURE 1 summarizes the complete workflow and emphasizes the separation between development-stage OOF processing and independent patient-level testing.

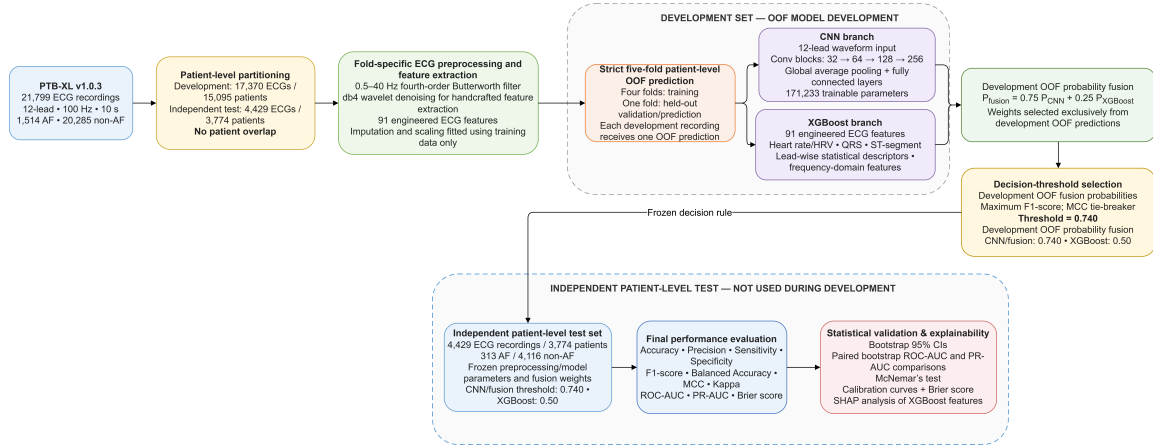


Figure 1: Overall workflow of the proposed patient-level OOF-guided threshold-aware heterogeneous CNN–XGBoost probability-fusion framework for atrial fibrillation detection.

2.2 PTB-XL Dataset

The experiments used the PTB-XL version 1.0.3 database [23]. After application of the study-specific labeling criterion, 21,799 twelve-lead ECG recordings sampled at 100 Hz, each with a duration of 10s, were incorporated in the analysis. AF cases were identified using the AFIB diagnostic annotation in the PTB-XL metadata. The resulting dataset contained 1,514 AF recordings and 20,285 non-AF recordings from 18,869 patients.

To prevent patient-level information from crossing the development–test boundary, the dataset was partitioned at the patient level before model development. The development partition contained 17,370 ECG recordings from 15,095 patients, including 1,201 AF recordings and 16,169 non-AF recordings. The independent test partition contained 4,429 ECG recordings from 3,774 patients, including 313 AF recordings and 4,116 non-AF recordings. No patient was shared between the development and independent test partitions.

All model selection, data-derived feature-processing steps, fusion-weight selection, and decision-threshold selection were performed using the development data. The independent test partition was retained entirely for the final assessment of the finalized models and was not used during model development or selection. The dataset characteristics and patient-level partitioning used in the experiments are summarized in TABLE 1.

Table 1: Characteristics of the PTB-XL dataset and patient-level experimental partitioning used for model development and independent evaluation.

Dataset Characteristic	Value
Total ECG recordings	21,799
Total patients	18,869
AF recordings	1,514 (6.95%)
Non-AF recordings	20,285 (93.05%)
Development ECG recordings	17,370
Development patients	15,095
AF recordings in development set	1,201
Non-AF recordings in development set	16,169
Independent test ECG recordings	4,429
Independent test patients	3,774
AF recordings in independent test set	313
Non-AF recordings in independent test set	4,116
Number of ECG leads	12
Sampling frequency	100 Hz
Recording duration	10 s
Handcrafted ECG features	91

2.3 ECG Preprocessing and Feature Extraction

ECG recordings were represented as 10-s, 12-lead signals sampled at 100 Hz, corresponding to 1000 samples per lead. For the handcrafted-feature branch, the ECG signals were converted to floating-point representation, centered using the median of each lead, and filtered using a fourth-order Butterworth band-pass filter with cutoff frequencies of 0.5 and 40 Hz. Wavelet denoising based on the Daubechies-4 (db4) wavelet was subsequently applied before extraction of the handcrafted features. No 50-Hz notch filter was applied. The handcrafted representation comprised 91 ECG-derived features. Seven global features were calculated from Lead II, including mean heart rate, SDNN, RMSSD, pNN50, RR-interval count, QRS width, and median ST-segment deviation [25]. The remaining 84 features consisted of seven descriptors calculated independently for each of the 12 leads: mean, standard deviation, root-mean-square amplitude, signal energy, and power in the 0–4, 4–8, and 8–15 Hz frequency bands. Thus, the feature representation contained 7 global and 84 lead-wise features.

QRS-related features could be unavailable when the implemented R-peak/QRS detection procedure did not identify a valid QRS-width estimate. In the development-set OOF experiments, missing QRS-width values were assigned using the median estimated from the corresponding training fold. Feature scaling was implemented using a StandardScaler fitted on the corresponding training fold and subsequently applied to its held-out fold. The independent test-set feature processing used imputation and scaling parameters fitted on the complete development set only. For the CNN branch, the ECG signals were processed separately using the fixed 0.5–40 Hz fourth-order Butterworth band-pass filter followed by per-lead median/MAD normalization. The resulting CNN input had a dimension of 1000×12 . The CNN preprocessing did not use the db4 wavelet-denoising step used

in the handcrafted-feature branch. No parameters for preprocessing, imputation, or scaling were estimated from the independent test set.

2.4 Classification Models

Decision Tree (DT), Random Forest (RF), AdaBoost, Extra Trees (ET), and XGBoost were evaluated using the 91 engineered ECG features. These models were included to compare conventional and ensemble-based classifiers within the engineered-feature representation. The waveform-based branch used a convolutional neural network (CNN) trained directly on the twelve-lead ECG waveforms. The CNN comprised four convolutional blocks with 32, 64, 128, and 256 filters, followed by global average pooling and fully connected layers. The network contained 171,233 trainable parameters and was trained using binary cross-entropy loss with the Adam optimizer. XGBoost was selected as the engineered-feature branch for the proposed heterogeneous probability-fusion framework, while the CNN constituted the waveform-based branch. The configurations of all evaluated models are summarized in TABLE 2.

Table 2: Model configuration of the machine learning models evaluated in this study.

Model	Hyperparameter	Value
Decision Tree (DT)	Criterion	Gini
	Random state	42
Random Forest (RF)	Number of trees	300
	Criterion	Gini
	Random state	42
AdaBoost (AB)	Number of estimators	200
	Random state	42
Extra Trees (ET)	Number of trees	300
	Criterion	Gini
	Random state	42
XGBoost	Number of estimators	600
	Maximum depth	6
	Learning rate	0.03
	Random state	42
CNN	Input Size	1000×12
	Convolutional filters	32, 64, 128, 256
	Dropout rate	0.40
	Batch size	64
	Learning rate	0.001
	Maximum epochs	25
	Trainable parameters	171,233

2.5 CNN–XGBoost Probability Fusion and Threshold Selection

The CNN and XGBoost models generated probability estimates of AF for each ECG recording. These probabilities were integrated using a weighted probability-fusion strategy based exclusively on development-set out-of-fold (OOF) predictions.

For an ECG recording i , the fused probability was calculated as

$$P_{fusion} = 0.75P_{CNN} + 0.25P_{XGBoost}$$

where P_{CNN} and $P_{XGBoost}$ denote the predicted probabilities of AF from the CNN and XGBoost models, respectively. Fusion weights were selected from the development-set OOF predictions by evaluating CNN weights from 0.00 to 1.00 in increments of 0.05, with the corresponding XGBoost weight defined as 1 minus the CNN weight. The configuration maximizing OOF ROC-AUC was selected, yielding a CNN weight of 0.75 and an XGBoost weight of 0.25. These weights were fixed before evaluation on the independent patient-level test set. The decision threshold for converting the fused probability into a binary AF classification was also selected exclusively from the development-set OOF fusion probabilities. The threshold was chosen by maximizing the F1-score, with Matthews correlation coefficient (MCC) used as the tie-breaking criterion. This procedure yielded a threshold of 0.740, which was fixed before independent test evaluation. Thus, both the fusion weights and the final decision threshold were determined entirely from development-stage OOF predictions, while the independent patient-level test set remained untouched until final evaluation.

2.6 Performance Evaluation and Statistical Validation

Model performance was assessed using accuracy, precision, sensitivity, specificity, F1-score, balanced accuracy, Matthews correlation coefficient (MCC), Cohen’s kappa, ROC-AUC, PR-AUC, and Brier score. Strict patient-level five-fold out-of-fold (OOF) prediction was performed within the development set. In each fold, four patient-level folds were used for model training and the remaining patient-level fold was held out for prediction. The CNN and XGBoost models were independently trained on the four training folds, and predictions were generated for the corresponding held-out fold. This procedure was repeated across all five folds so that every development-set recording received exactly one OOF prediction from a model that had not been trained using recordings from the same patient.

Fold-specific data-derived preprocessing parameters, feature imputation, and feature scaling were determined using training data only and then applied to the corresponding held-out fold. CNN preprocessing used the predefined filtering and normalization procedure described in Section 2.3. The resulting OOF probabilities were used for model comparison, fusion-weight selection, and decision-threshold selection. The independent patient-level test set was not used during OOF prediction, fusion-weight selection, or threshold optimization. After model development and selection were completed, the finalized models, fusion weights, and decision threshold were evaluated on the independent patient-level test set. Bootstrap resampling of the independent test predictions was used to estimate 95% confidence intervals for the principal performance measures.

Paired bootstrap comparisons were performed between the proposed fusion and each individual model to estimate differences in ROC-AUC and PR-AUC, together with corresponding 95% confidence intervals and p-values. McNemar’s test was used to compare paired classification outcomes at the specified fixed decision thresholds. Calibration was assessed using the Brier score and calibration curves [24]. The Brier score was used as a quantitative measure of probabilistic prediction error, with lower values indicating lower squared probability error on the evaluated test set. SHAP analysis was used to examine the contribution of engineered ECG features in the XGBoost component [26]. For fixed operating-point comparison, the development-selected threshold of 0.740 was applied to the CNN probabilities, whereas XGBoost was evaluated using its fixed threshold of 0.50.

3. RESULTS AND DISCUSSION

3.1 Development OOF Results and Fusion Selection

Strict patient-level five-fold OOF prediction was performed within the development set, with each recording receiving a prediction from a model that had not been trained using recordings from the same patient. The CNN achieved an OOF ROC-AUC of 0.9907 and PR-AUC of 0.8938, whereas XGBoost achieved an OOF ROC-AUC of 0.9628 and PR-AUC of 0.7268.

The OOF probabilities from the two branches were used to evaluate the probability-fusion configuration without using the independent test set. Combining CNN and XGBoost probabilities using weights of 0.75 and 0.25, respectively, produced an OOF ROC-AUC of 0.9919 and PR-AUC of 0.9146. At the reference threshold of 0.50, the fusion yielded an F1-score of 0.8448 and MCC of 0.8381.

The final decision threshold was subsequently selected from the development-set OOF fusion probabilities by maximizing F1-score, with MCC used as the tie-breaking criterion. This procedure yielded a threshold of 0.740. Neither the labels nor the predictions from the independent patient-level test set were used during threshold selection, and the selected threshold was fixed before independent test evaluation.

TABLE 3 presents the development-set OOF performance and fusion results.

Table 3: Development-set out-of-fold performance of the evaluated models and fusion strategies.

Model/Configuration	OOF ROC-AUC	OOF PR-AUC	F1-score	MCC
XGBoost OOF	0.9628	0.7268	0.6146	0.6090
CNN OOF	0.9907	0.8938	0.8284	0.8226
CNN–XGBoost Fusion	0.9919	0.9146	0.8448	0.8381

Note: F1-score and MCC in this table are reported at a threshold of 0.50. The final decision threshold of 0.740 was selected separately from the development-set OOF fusion probabilities by maximizing F1-score, with MCC as the tie-breaking criterion, and was fixed before independent test evaluation.

3.2 Independent Test Performance

After completion of development-stage model selection, fusion-weight selection, and decision-threshold selection, the finalized models were evaluated on the independent patient-level test set comprising 4,429 ECG recordings from 3,774 patients, including 313 AF and 4,116 non-AF recordings. No patient was shared between the development and independent test partitions, and the test set was not used for model training, fusion-weight selection, or threshold optimization. The performance of the evaluated classifiers is presented in TABLE 4. XGBoost achieved a ROC-AUC of 0.9575 and PR-AUC of 0.7413, with an accuracy of 95.46% and F1-score of 60.97%. The CNN achieved a ROC-AUC of 0.9892, PR-AUC of 0.9113, accuracy of 98.37%, and F1-score of 88.92%.

Table 4: Performance of the evaluated classifiers on the independent patient-level test set.

Model	Threshold	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1 (%)	BA (%)	MCC	κ	ROC-AUC (%)	PR-AUC (%)	Brier
Decision Tree	0.50	91.35	39.88	44.09	94.95	41.88	69.52	0.373	0.372	69.75	24.11	0.0842
Random Forest	0.50	94.04	81.01	20.45	99.64	32.65	60.04	0.389	0.307	94.23	61.85	0.0425
AdaBoost	0.50	92.93	0.00	0.00	100.00	0.00	50.00	0.000	0.000	84.87	35.63	0.0910
Extra Trees	0.50	94.02	82.43	19.49	99.68	31.52	59.59	0.383	0.296	93.92	63.33	0.0438
XGBoost	0.50	95.46	77.72	50.16	98.91	60.97	74.53	0.603	0.587	95.75	74.13	0.0335
CNN	0.740	98.37	85.76	92.33	98.83	88.92	95.58	0.881	0.880	98.92	91.13	0.0164
CNN–XGBoost Fusion	0.740	98.56	90.29	89.14	99.27	89.71	94.20	0.889	0.889	98.97	94.53	0.0134

Note: BA, balanced accuracy; MCC, Matthews correlation coefficient; κ , Cohen’s kappa. ROC-AUC and PR-AUC are threshold-independent. For the CNN, 0.740 denotes the development-selected fusion threshold applied for fixed operating-point comparison; XGBoost was evaluated at its fixed threshold of 0.50.

The proposed CNN–XGBoost fusion achieved a ROC-AUC of 0.9897 and PR-AUC of 0.9453. At the development-selected threshold of 0.740, the fusion achieved 98.56% accuracy, 90.29% precision, 89.14% sensitivity, 99.27% specificity, and 89.71% F1-score. Balanced accuracy was 94.20%, MCC was 0.889, and Cohen’s kappa was 0.889. The Brier score was 0.0134. Relative to the CNN, the fusion increased PR-AUC from 0.9113 to 0.9453, while ROC-AUC changed from 0.9892 to 0.9897. The corresponding statistical comparison of these differences is presented in Section 3.3. The ROC curves of the evaluated classifiers on the independent patient-level test set are shown in FIGURE 2. The precision–recall curves for the evaluated classifiers on the independent patient-level test set are presented in FIGURE 3.

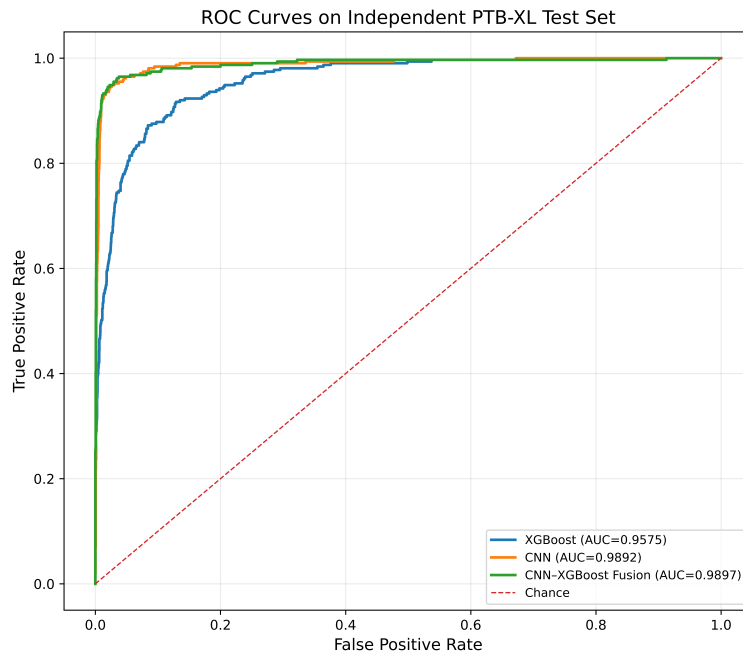


Figure 2: Receiver operating characteristic curves of the evaluated classifiers on the independent patient-level test set.

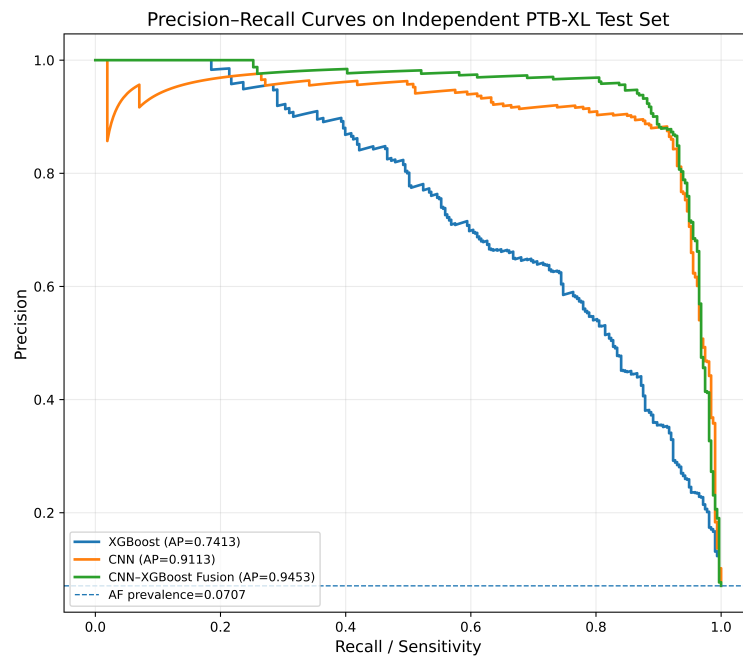


Figure 3: Precision-recall curves of the evaluated classifiers on the independent patient-level test set.

3.3 Statistical Validation and Calibration

Bootstrap resampling of the independent patient-level test predictions was used to estimate 95% confidence intervals for the principal performance measures. The proposed fusion achieved a ROC-AUC of 0.9897 (95% CI: 0.9820–0.9953), PR-AUC of 0.9453 (95% CI: 0.9221–0.9640), accuracy of 98.56% (95% CI: 98.22–98.89%), F1-score of 89.71% (95% CI: 87.14–91.99%), and MCC of 0.889 (95% CI: 0.861–0.914). The corresponding confidence intervals are shown in FIGURE 4. The bootstrap 95% confidence intervals for the principal performance metrics and the Brier scores of all evaluated models are reported in TABLE 5. Paired bootstrap comparisons showed that the difference in ROC-AUC between the fusion and CNN was 0.0004 (95% CI: -0.0033 to 0.0042; $p = 0.789$), indicating no statistically supported difference in ROC-AUC. In contrast, the fusion showed a PR-AUC increase of 0.0331 relative to the CNN (95% CI: 0.0051–0.0676; $p = 0.018$). Compared with XGBoost, the fusion showed differences of 0.0322 in ROC-AUC (95% CI: 0.0245–0.0403; $p < 0.001$) and 0.2039 in PR-AUC (95% CI: 0.1659–0.2442; $p < 0.001$).

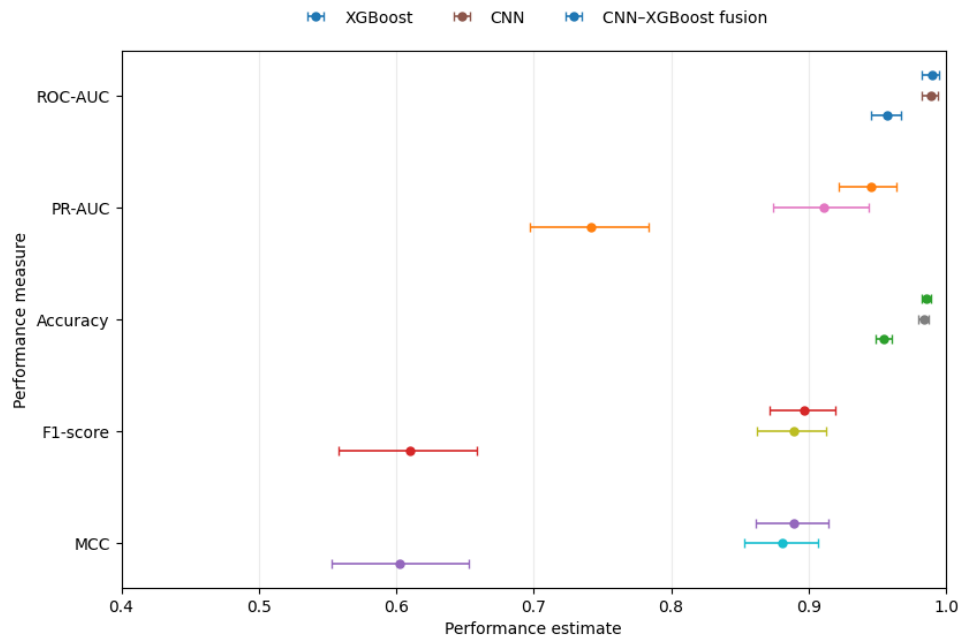


Figure 4: Bootstrap 95% confidence intervals for the principal performance measures of XGBoost, CNN, and the proposed fusion.

McNemar's test yielded $p = 0.185$ for the fusion versus CNN comparison and $p < 0.001$ for the fusion versus XGBoost comparison at their respective fixed operating thresholds. Calibration was assessed using Brier scores and calibration curves. The Brier scores were 0.0335 for XGBoost, 0.0164 for CNN, and 0.0134 for the fusion, with the corresponding calibration curves shown in FIGURE 5.

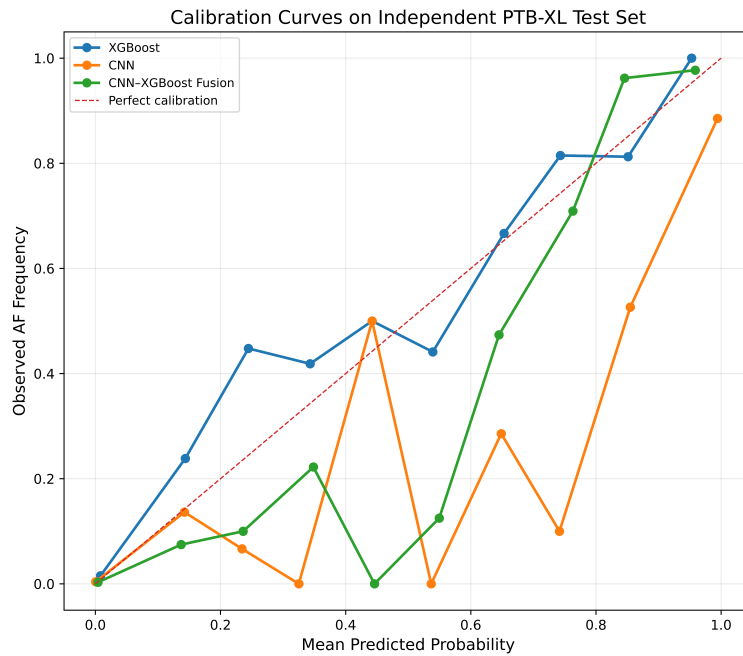


Figure 5: Calibration curves for XGBoost, CNN, and the proposed CNN–XGBoost fusion on the independent test set.

Table 5: Bootstrap 95% confidence intervals and Brier scores for the evaluated models on the independent patient-level test set.

Model	ROC-AUC (95% CI)	PR-AUC (95% CI)	Accuracy (95% CI)	F1-score (95% CI)	MCC (95% CI)	Brier
XGBoost	0.9575 (0.9455–0.9670)	0.7413 (0.6976–0.7837)	95.46 (94.85–96.05)	60.97 (55.78–65.83)	0.603 (0.553–0.653)	0.0335
CNN	0.9892 (0.9825–0.9942)	0.9113 (0.8744–0.9440)	98.37 (98.01–98.74)	88.92 (86.26–91.30)	0.881 (0.853–0.907)	0.0164
CNN– XGBoost Fusion	0.9897 (0.9820–0.9953)	0.9453 (0.9221–0.9640)	98.56 (98.22–98.89)	89.71 (87.14–91.99)	0.889 (0.861–0.914)	0.0134

3.4 Explainability of the Feature-Based Component

SHAP analysis was implemented for the XGBoost branch to examine the influence of the engineered ECG features. The SHAP summary plot showing the contribution and direction of these features is presented in FIGURE 6. The analysis identified contributions from heart-rate and heart-rate-variability measures, QRS-related characteristics, ST-segment measures, lead-wise statistical descriptors, and frequency-domain features. These findings provide an interpretable view of the engineered-feature branch and complement the waveform-based representation learned by the CNN.

Because SHAP was applied only to XGBoost, the resulting feature contributions should not be taken as explanations of the CNN representations or of the complete CNN–XGBoost fusion model.

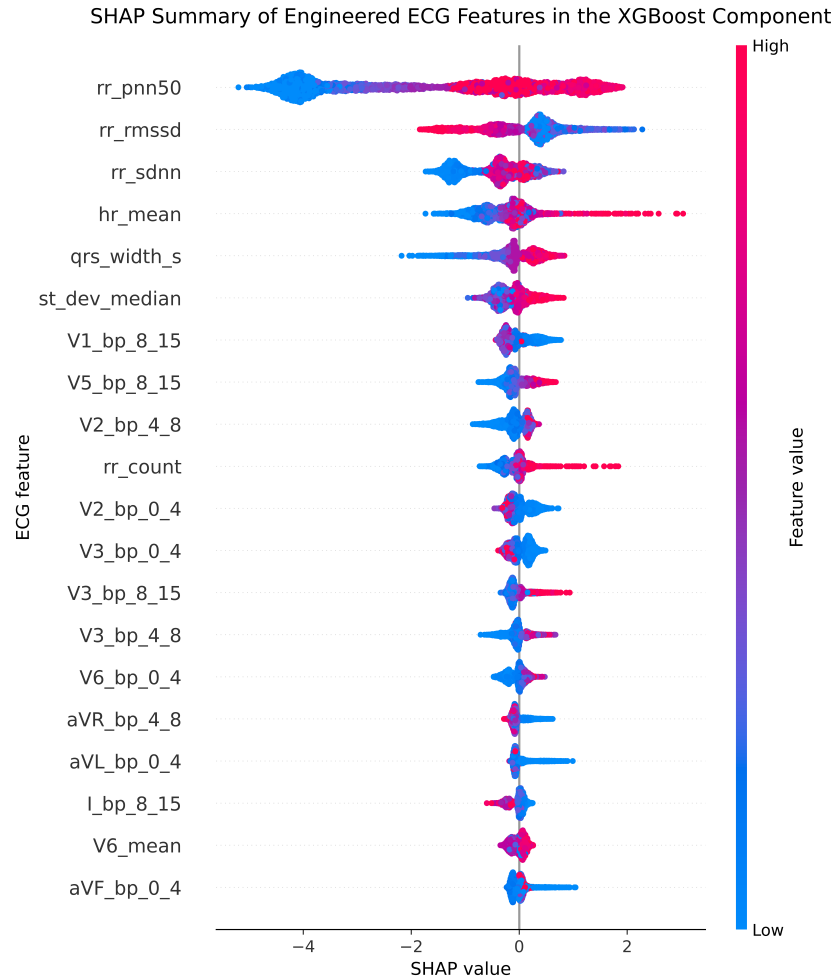


Figure 6: SHAP summary plot showing the contribution and direction of engineered ECG features in the XGBoost component.

3.5 Discussion

The results indicate that the CNN and XGBoost branches capture complementary representations of ECG information. The CNN learns directly from twelve-lead waveforms, whereas XGBoost operates on engineered temporal, morphological, statistical, and frequency-domain features. The proposed probability fusion therefore combines two distinct representations rather than relying on a single input representation. On the independent test set, the fusion increased PR-AUC from 0.9113 for the CNN to 0.9453 and reduced the Brier score from 0.0164 to 0.0134. These changes indicate that the principal observed benefit of the heterogeneous fusion was in precision–recall performance and probabilistic prediction error rather than a material change in ROC-AUC.

The statistical analysis provides an important qualification to this observation. The difference in ROC-AUC between the fusion and CNN was 0.0004 and was not statistically supported ($p = 0.789$), whereas the increase in PR-AUC of 0.0331 was statistically supported ($p = 0.018$). Compared with XGBoost, statistically supported differences were observed for both ROC-AUC and PR-AUC. Thus, the results support the value of evaluating the heterogeneous fusion using metrics beyond accuracy and ROC-AUC, particularly in the presence of substantial class imbalance.

An important practical aspect of the study is the separation of model development from final evaluation. Patient-level partitioning prevented recordings from the same patient from appearing across the development and independent test sets. Within the development set, five-fold OOF predictions were used for fusion-weight selection and decision-threshold selection, while the independent test set remained untouched until final evaluation. The decision threshold of 0.740 was therefore determined without using independent test predictions. This design reduces the risk of information leakage during model selection and provides a clearer estimate of performance on previously unseen patients.

The heterogeneous fusion also introduces additional modelling complexity compared with using the CNN alone. The observed increase in PR-AUC and reduction in Brier score provide empirical justification for evaluating this added component, but the computational and implementation costs of maintaining both branches should be considered in practical deployment. The present results therefore support the methodological utility of the fusion within the evaluated setting rather than establishing that the additional branch is necessary for every AF-detection application.

There are several limitations to this study. First, the analysis was conducted using PTB-XL, and external validation on independent multi-centre and prospective ECG datasets is required to assess generalizability across patient populations, acquisition systems, and clinical settings [27]. Second, the bootstrap analysis was performed at the ECG-recording level and did not explicitly account for within-patient clustering. Third, SHAP analysis was limited to the engineered-feature/XGBoost component and therefore does not provide a complete explanation of the waveform-based CNN or the fused prediction. Finally, the fixed fusion weights and decision threshold were selected within the development dataset; their robustness should be evaluated in external datasets before clinical use.

4. CONCLUSION

This study presented a patient-level heterogeneous CNN–XGBoost probability-fusion framework for atrial fibrillation detection using the PTB-XL database. The framework combines waveform representations learned from twelve-lead ECGs with engineered ECG features and uses development-set out-of-fold predictions to determine the fusion configuration and decision threshold while preserving an independent patient-level test set for final evaluation. On the independent test set, the proposed fusion achieved an accuracy of 98.56%, F1-score of 89.71%, ROC-AUC of 0.9897, PR-AUC of 0.9453, and Brier score of 0.0134 at the fixed decision threshold of 0.740. Compared with the CNN, the fusion produced a statistically supported increase in PR-AUC, whereas the difference in ROC-AUC was not statistically supported. These findings indicate that the principal observed benefit of combining the two representations was in precision–recall performance and probabilistic prediction error. The study also demonstrates the importance of patient-level separa-

tion and development-only OOF procedures when evaluating ensemble ECG models. However, the findings are based on PTB-XL, and external validation on independent multi-centre and prospective ECG datasets is required before assessing generalizability to other populations and clinical settings. Future work should also evaluate the robustness of the fusion configuration and threshold across external datasets and account explicitly for patient-level clustering in statistical validation.

References

- [1] Hindricks G, Potpara T, Dagres N, Arbelo E, Bax JJ, et al. 2020 ESC Guidelines for the Diagnosis and Management of Atrial Fibrillation Developed in Collaboration With the European Association for Cardio-Thoracic Surgery (EACTS) the Task Force for the Diagnosis and Management of Atrial Fibrillation of the European Society of Cardiology (ESC) Developed With the Special Contribution of the European Heart Rhythm Association (EHRA) of the ESC. *European Heart Journal*. 2021;42:373-498. <https://doi.org/10.1093/eurheartj/ehaa612>
- [2] Ma C, Xiao Z, Zhao L, Biton S, Behar JA, et al. A Review on Atrial Fibrillation Detection From Ambulatory ECG. *IEEE Transactions on Biomedical Engineering*. 2024;71:876-892. <https://doi.org/10.1109/TBME.2023.3321792>
- [3] Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep Learning for ECG Analysis: Benchmarks and Insights From PTB-XL. *IEEE journal of biomedical and health informatics*. 2021;25:1519-1528. <https://doi.org/10.1109/JBHI.2020.3022989>
- [4] Xie C, Wang Z, Yang C, Liu J, Liang H. Machine Learning for Detecting Atrial Fibrillation From ECGs: Systematic Review and Meta-Analysis. *Reviews in cardiovascular medicine*. 2024;25:8. <https://doi.org/10.31083/j.rcm2501008>
- [5] Choi S, Choi K, Yun HK, Kim SH, Choi HH, et al. Diagnosis of Atrial Fibrillation Based on Ai-Detected Anomalies of ECG Segments. *Heliyon*. 2024 Jan 15;10(1). <https://doi.org/10.1016/j.heliyon.2023.e23597>
- [6] Yang MU, Lee DI, Park S. Automated Diagnosis of Atrial Fibrillation Using ECG Component-Aware Transformer. *Computers in Biology and Medicine*. 2022;150:106115. <https://doi.org/10.1016/j.compbimed.2022.106115>
- [7] Lim J, Han D, Chon KH. QRS-Centric Beat-Wise Atrial Fibrillation Detection in ECG Signals Using Deep Neural Networks. *Computers in biology and medicine*. 2025;192:110282. <https://doi.org/10.1016/j.compbimed.2025.110282>
- [8] Khaja S, Baijoo K, Aziz R. Artificial Intelligence-Powered Advancements in Atrial Fibrillation Diagnostics: A Systematic Review. *The Egyptian Heart Journal*. 2025;77:73. <https://doi.org/10.1186/s43044-025-00670-y>
- [9] Lin WY, Lin C, Liu WC, Liu WT, Chang CH, et al. Development of an Artificial Intelligence-Enabled Electrocardiography to Detect 23 Cardiac Arrhythmias and Predict Cardiovascular Outcomes. *Journal of medical systems*. 2025;49:51. <https://doi.org/10.1007/s10916-025-02177-0>
- [10] Liu Q, Zhou Y, He J, Hou J, Zhang D, et al. AF-ECGNET: An Effective Model for Atrial Fibrillation Intelligent Detection Based on the Improved Transformer Model. *Biomedical Signal Processing and Control*. 2026;116:109493. <https://doi.org/10.1016/j.bspc.2026.109493>

- [11] Kraft D, Rumm P. Atrial Fibrillation and Atrial Flutter Detection Using Deep Learning. *Sensors*. 2025;25:4109. <https://doi.org/10.3390/s25134109>
- [12] De Guio F, Rienstra M, Lillo-Castellano JM, Toribio-Fernández R, Lizcano C, et al. Enhanced Detection of Atrial Fibrillation in Single-Lead Electrocardiograms Using a Cloud-Based Artificial Intelligence Platform. *Heart Rhythm*. 2025;22:1667-1674. <https://doi.org/10.1016/j.hrthm.2024.12.048>
- [13] Atwa AE, Atlam ES, Ahmed A, Atwa MA, Abdelrahim EM, et al. Interpretable Deep Learning Models for Arrhythmia Classification Based on ECG Signals Using PTB-X Dataset. *Diagnostics*. 2025;15:1950. <https://doi.org/10.3390/diagnostics15151950>
- [14] Śmigiel S, Pałczyński K, Ledziński D. Deep Learning Techniques in the Classification of ECG Signals Using R-Peak Detection Based on the PTB-XL Dataset. *Sensors*. 2021;21:8174. <https://doi.org/10.3390/s21248174>
- [15] Pałczyński K, Śmigiel S, Ledziński D, Bujnowski S. Study of the Few-Shot Learning for ECG Classification Based on the PTB-XL Dataset. *Sensors*. 2022;22:904. <https://doi.org/10.3390/s22030904>
- [16] Phukan N, Manikandan MS, Pachori RB. Fast and Resource Efficient Atrial Fibrillation Detection Framework for Long Term Health Monitoring Devices. *IEEE Sensors Letters*. 2024;8:1-4. <https://doi.org/10.1109/LSENS.2024.3367724>
- [17] Jin Y, Ko B, Chang W, Choi KH, Lee KH. Explainable Paroxysmal Atrial Fibrillation Diagnosis Using an Artificial Intelligence-Enabled Electrocardiogram. *The Korean journal of internal medicine*. 2025;40:251. <https://doi.org/10.3904/kjim.2024.130>
- [18] Jo YY, Cho Y, Lee SY, Kwon JM, Kim KH, et al. Explainable Artificial Intelligence to Detect Atrial Fibrillation Using Electrocardiogram. *International journal of cardiology*. 2021;328:104-110. <https://doi.org/10.1016/j.ijcard.2020.11.053>
- [19] Yang Z, Jin A, Li Y, Yu X, Xu X, et al. A Coordinated Adaptive Multiscale Enhanced Spatio-Temporal Fusion Network for Multi-Lead Electrocardiogram Arrhythmia Detection. *Scientific reports*. 2024;14:20828. <https://doi.org/10.1038/s41598-024-71700-z>
- [20] Baj G, Gandin I, Scagnetto A, Bortolussi L, Cappelletto C, et al. Comparison of Discrimination and Calibration Performance of Ecg-Based Machine Learning Models for Prediction of New-Onset Atrial Fibrillation. *BMC Medical Research Methodology*. 2023;23:169. <https://doi.org/10.1186/s12874-023-01989-3>
- [21] Parker G, Phan J, Lovell N, Subbiah RN, Vandenberg JI, et al. MultiECGNet: Evaluation of a Deep Learning Ensemble Method for Image-Based 12-Lead Electrocardiographic Detection of Atrial Fibrillation. *CJC Open*. 2026;8:p1074-1087 <https://doi.org/10.1016/j.cjco.2026.03.022>
- [22] Habib M, Bikai M, Merhi Z, Rammal R, El Masri A. Calibration-Aware Hybrid Machine Learning and Deep Learning Ensemble for Cardiovascular Disease Prediction. *Adv Artif Intell Mach Learn*. 2026. (Ahead of Print). <https://dx.doi.org/10.54364/AAIML.2026.65341>
- [23] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, et al. PTB-XL, A Large Publicly Available Electrocardiography Dataset. *Scientific data*. 2020;7:154. <https://doi.org/10.1038/s41597-020-0495-6>

- [24] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17:230. <https://doi.org/10.1186/s12916-019-1466-7>
- [25] Rawshani A, Rawshani A, Smith G, Boren J, Bhatt DL, et al. Integrating Deep Learning With Ecg, Heart Rate Variability and Demographic Data for Improved Detection of Atrial Fibrillation. *Open Heart.* 2025;12:e003185. <https://doi.org/10.1136/openhrt-2025-003185>
- [26] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, et al. From Local Explanations to Global Understanding With Explainable AI for Trees. *Nature machine intelligence.* 2020;2:56-67. <https://doi.org/10.1038/s42256-019-0138-9>
- [27] Lee KJ, Seo CO, Jeong JH, Kim Y, Lee KY, et al. A Multicenter, Randomized, Controlled Study to Evaluate the Efficacy of an Artificial Intelligence-Based Electrocardiography for the Detection of Atrial Fibrillation in Patients With Suspected Symptoms (DEEP-AF). *Heart Rhythm O₂.* 2026. <https://doi.org/10.1016/j.hroo.2026.07.009>