Divergence-Optimized Data Fusion that Enhances AI-Based Electrocardiogram Arrhythmia Detection

Baozhuo Su[1,4], Qingli Dou[2,4], Kang Liu[1,4], Zhengxian Qu[1], Jerry Deng[1], Ting Tan[1,5], Zhiqun Tan[3,5], and Yanan Gu[2,5]

[1]Aerie Intelligent Technology Corp, Irvine, CA, USA
[2]The People's Hospital of Baoan Shenzhen, Guangdong, China
[3]Department of Anatomy and Neurobiology, University of California Irvine School of Medicine, Irvine, CA92697
[4]These authors contribute equally.
[5]Corresponding authors' email: tting@aerieintech.com (T.T.), tanz@hs.uci.edu (Z.T.), gyn1252006@163.com (Y.G.)

ABSTRACT

Accurate arrhythmia detection using artificial intelligence remains a significant challenge, particularly when training data are limited in size and heterogeneous in origin. To address this issue, we present a novel data fusion strategy that leverages gradient descent optimization to minimize the Kullback–Leibler divergence between feature distributions across datasets from different sources. We evaluate this approach using a combination of in-house electrocardiogram (ECG) recordings collected from both healthy and diseased individuals using various machines and a subset of data extracted from the publicly available PTB-XL dataset. Binary classification is performed using Light Gradient Boosting Machine (LightGBM) and eXtreme Gradient Boosting (XGBoost) models. Applying the fusion strategy leads to consistently higher classification accuracy on both in-house and PTB-XL datasets compared to standard normalization methods. These results suggest that aligning feature distributions across datasets can improve arrhythmia classification and offer a generalizable strategy for integrating heterogeneous clinical data in machine-learning models.

Introduction

When AI computation is used in the field of healthcare, challenges are particularly posed with small heterogeneous clinical datasets, as large volumes of high-quality data are required to train effective

models in the context of machine learning. Thus, integrating clinical datasets from diverse sources is a crucial step toward enabling robust and efficient AI computation. However, this integration is complicated by the inherent heterogeneity, complexity, and limited representativeness of many clinical datasets[1].

In recent years, the application of AI to analysis of electrocardiogram (ECG) data for the detection and classification of arrhythmias has emerged as a highly active and extensively studied area in clinical research[2,3]. These methods typically involve training classical machine learning or deep learning models using extracted features and corresponding labels from annotated datasets. Among the various available learning architectures, gradient boosting models are particularly valued for their strong explainability, providing transparent decision paths and feature importance metrics, which are essential in medical applications. Light Gradient Boosting Machine (LightGBM) and eXtreme Gradient Boosting (XGBoost) are two of the most effective and interpretable implementations[4,5]. Beyond model selection, the quality, quantity, and consistency of the training data play a critical role in determining overall model performance. When utilizing data from multiple sources, the aforementioned challenge of effective data integration becomes especially important and must be carefully addressed.

To address this data-fusion issue, many different approaches have recently been proposed to improve the effectiveness of AI computation in medical contexts. A specialized data quality framework, known as the METRIC framework, was recently proposed by Schwabe and colleagues[6]. Through evaluating the diversity of datasets across demographic factors, data sources, and accuracy of measurements relevant to both device and human errors, and so forth, the METRIC framework offered a systematic approach to evaluating data quality and reducing biases of small clinical datasets. However, this framework did not provide standardized guidelines for measuring and evaluating data quality. This limitation was then addressed through the adoption of a standardized dataset based on the Fast Healthcare Interoperability Resources (FHIR) standard, which enhances data interoperability and

facilitates sharing across various healthcare systems[7]. In the context of ECG data, traditional data fusion and preprocessing techniques (such as amplitude normalization, band-pass filtering, and period scaling) are commonly used to reduce noise and standardize input[2,8]. While these methods can improve signal clarity, however, they may inadvertently remove clinically relevant information, potentially compromising diagnostic accuracy.

A more suitable perspective for addressing data integration challenges originates from the field of statistics. Kernel density estimation (KDE) and Kullback–Leibler (KL) divergence are widely used statistical tools in machine learning and information theory, with established applications in anomaly and outlier detection, for example, identifying regions of unusually low data density in cybersecurity or industrial process monitoring. Previous research in signal processing and machine learning demonstrated that KDE and KL divergence could enhance the performance and robustness of algorithms by effectively reducing redundant information between different distributions, thereby improving both efficiency and accuracy[9,10]. While KDE and KL divergence have been successfully applied in fields such as robotics and data exploration[11,12], their potential remains underexplored in medical signal processing. In particular, they have not yet been leveraged to estimate feature distributions in clinical data, which could offer new opportunities to improve data integration and representation in healthcare AI applications.

In this study, we propose a novel method that integrates KL divergence optimization of feature distributions estimated from KDE with machine learning techniques to merge ECG feature datasets when performing joint analysis, aiming to enhance detection performance while minimizing computational cost. Importantly, we bifurcated the processing of Gaussian and non-Gaussian feature distributions in the datasets. Features with Gaussian distributions were fused based on their mean and variance. For features with non-Gaussian distributions, we first quantified the distributional discrepancy between the two datasets using KL divergence Specifically, we apply Gaussian kernel–based distribution estimation to identify inter-dataset parameters with linear relationships and minimize their KL divergence, ultimately

achieving a reduced KL divergence. Given that LightGBM and XGBoost are among the most commonly used models in clinical ECG analysis, the resulting data are subsequently evaluated by these two models. This approach with the bifurcated treatment of Gaussian and non-Gaussian features leverages traditional and modern machine learning techniques to enhance the classification accuracy and robustness of arrhythmia detection systems by effectively reducing redundancy and errors in multi-source datasets.

The main structure of our paper can be divided into Results, Discussions, and Datasets and Methods sections according to the editorial instructions of the Journal. The Results section presents feature distributions, detailed model performance metrics including confusion matrices and Receiver Operating Characteristic (ROC) curves, and computation outcomes for each experimental case. The Discussion section provides an in-depth analysis of the results obtained from our methods, explores their potential application scenarios, and outlines possible future improvements. Finally, a comprehensive description of the data sources, data collection and processing procedures, and the implementation of the algorithms are described in the Datasets and Methods section.

RESULTS

To test our fusion algorithms, we recruit a total of seven different datasets, i.e., three in-house datasets directly collected from clinical patients and four datasets derived from PTB-XL ECG database. Tables S1 and S2 (Supplementary Information) summarize the data counts and relevant details for each dataset. We also construct additional datasets by combining in-house and PTB-XL-derived data to further validate the fusion pipeline. As illustrated in Figure 1, the analysis process begins with sequential preprocessing steps applied to all ECG waveform data prior to model training. We extract a total of nine ECG signal features, with six showing Gaussian distribution patterns (e.g., approximate entropy, detrended fractal dimension, Higuchi fractal dimension, Katz fractal dimension, sample entropy, and spectral entropy) and three displaying non-Gaussian characteristics (e.g., Petrosian fractal dimension, permutation entropy,

singular value decomposition entropy), for use in data fusion computation. Consequently, we first evaluate the fusion algorithms compare them with standard normalization using in-house clinical datasets, then extend the assessment to PTB-XL-derived datasets and their combination with the in-house data, and finally perform a comparative analysis of the results.

Optimized data fusion algorithms enhance machine learning-based ECG classification performance on in-house datasets

The in-house datasets consist of three ECG cohorts: one diseased group and two healthy control groups, with different recording lengths and collected at two distinct sampling frequencies. After successfully integrating the healthy datasets, we use the optimized features from the combined dataset and compare them with those of the diseased dataset to train machine learning LightGBM and XGBoost models. Figure 2 shows the distribution patterns of the nine features in the In-house datasets. For features following a Gaussian distribution, we leverage the property that such distributions are preserved under affine transformations. For non-Gaussian features, we minimize the KL divergence between the distributions of the same feature across different datasets. These features are then used to classify data instances as "Healthy" or "Diseased".

Table 1 summarizes the classification accuracy, false positive rate, and false negative rate for the tested features on the in-house datasets using the LightGBM model. As shown, the Gaussian feature approximate entropy and non-Gaussian feature Approximate Entropy-based fusion algorithms yield the highest classification accuracies [0.992 (95% CI: 0.987–0.997) and 0.982 (95% CI: 0.971–0.998), respectively]. Comparable improvements are observed when all nine Gaussian and non-Gaussian features are integrated and optimized via fusion algorithms. Similar trends are also observed with the XGBoost model following fusion optimization (Table S3 in the Supplementary Information). Notably, this fusion approach with bifurcated treatments of Gaussian and non-Gaussian features produces significantly lower

false positive and false negative rates compared to standard normalization across all three cases, along with enhanced classification accuracy. Similar improvements are also seen in the Random Forest and Support Vector Machine models (see Supplementary Tables S6 and S9).

Optimized data fusion algorithms improve LightGBM and XGBoost based classification performance on the PTB-XL-derived datasets

To further evaluate the effectiveness of our fusion algorithms in enhancing ECG classification performance using LightGBM and XGBoost models, we utilize four small ECG datasets extracted from PTB-XL database. Each dataset has a distinct recording format that mimics the structure of the in-house datasets. Following the same processing pipeline used on the in-house datasets, the same nine ECG signal features are extracted and demonstrated in Figure 3. As summarized in Table 2, fusion optimizations based on the non-Gaussian features of permutation entropy and Petrosian fractal dimensions, as well as their combined features, achieve the best classification performance in the LightGBM model compared with other feature sets. These observations are consistent with those obtained using the XGBoost model (Supplementary Table S4) relative to standard normalization. Although the improvements are less pronounced than those observed with LightGBM or XGBoost, moderate gains are evident in the results from the Random Forest and Support Vector Machine algorithms across all three cases (Supplementary Tables S7 and S10).

Assessment of Optimized Data Fusion on Merged PTB-XL and In-House Datasets

After testing the fusion algorithms on the PTB-XL-derived datasets, we also apply the same processes described above to evaluate the applicability of our fusion pipeline on a heterogeneous dataset combining both in-house and PTB-XL-derived ECG data from Healthy and Diseased groups. As shown in Table 3 and Supplementary Table S5, the non-Gaussian features based optimized fusions enhance both LightGBM and XGBoost mediated ECG classification performance relative to standard normalization,

though the improvement is smaller in the Random Forest and Support Vector Machine models (Supplementary Tables S8 and S11).

Taken together, Table 4 provides a consolidated summary of key performance metrics from the three experiments described above (see Figure S1 in the Supplementary Information for the confusion matrices). A comparison of the ROC curves across all experimental combinations reveals that the optimized fusion methods effectively enhance ECG classification performance for both LightGBM and XGBoost models, with LightGBM showing slightly better results.

DISCUSSION

In this study, we demonstrate the effectiveness of a KDE and KL divergence–based optimization method for integrating multiple heterogeneous ECG signal datasets based on their Gaussian and non-Gaussian distribution features for arrhythmia classification using LightGBM and XGBoost models. This approach substantially enhances classification performance compared to standard normalization techniques. The application of the KDE and KL divergence–based optimization method for analyzing multi-source datasets reveals several key insights. The fusion algorithms effectively aggregate heterogeneous data sources by distinguishing Gaussian from non-Gaussian distributions and leverage the strengths of both Gaussian and non-Gaussian feature transformations, ensuring minimal redundancy while preserving essential information for machine learning models. Notably, features such as permutation entropy and approximate entropy exhibit substantial gains in accuracy after data fusion. Both LightGBM and XGBoost models showed similar responses to this optimization, demonstrating robustness across different classifiers. The integration of PTB-XL data further validates the approach, as the results demonstrate notable improvements over conventional normalization-based fusion strategies. These findings suggest that the divergence-optimized fusion approach effectively mitigates the inconsistencies introduced by heterogeneous datasets.

By applying distinct mapping functions tailored to each distribution type, the study highlights how the algorithm's performance varies under different statistical assumptions. While the primary goal was to improve classification accuracy, the analysis also reveals that outcomes are highly sensitive to the selected feature set. In this regard, the features of permutation entropy and approximate entropy-based fusion yield the best performance on the in-house datasets, while the non-Gaussian features of permutation entropy and Petrosian fractal dimension-based fusion achieve better classification performance than the other feature sets on the PTB-XL-derived datasets. This discrepancy likely arises from their parameter-normalized, amplitude-invariant quantification of signal complexity, which makes them resilient to noise, baseline drift, and sampling rate variations across heterogeneous ECG recordings. Specifically, approximate entropy normalizes pattern match counts by the total number of comparisons and accommodates a broad range of similarity thresholds, producing stable complexity estimates under diverse recording conditions. Similarly, permutation entropy relies only on the ordinal ranking of short subsequences, making it invariant to monotonic transformations and capable of suppressing common artifacts without extensive preprocessing. In contrast, features such as fractal dimensions, spectral entropy, DFA exponents, and SVD-based metrics require careful tuning of parameters like window lengths, embedding dimensions, and frequency band definitions. This sensitivity often leads to inconsistent performance when applied directly to new datasets. These discrepancies suggest that dataset-specific factors, such as differences in acquisition protocols and preprocessing pipelines, can significantly affect the utility of various features.

Another key finding is the strong correlation between classification improvement and the degree of feature distribution alignment achieved through fusion. Notably, the optimized method consistently showed better performance than standard approaches when applied to features with non-Gaussian distributions, a trend that warrants further investigation. Features exhibiting larger distributional shifts before and after fusion tend to yield greater gains in accuracy, supporting the hypothesis that optimizing

distributional alignment is essential for effective multi-source data integration. It is also worth noting that this data fusion strategy differs from recently developed domain adaptation approaches[13], which typically involve training on source data (one domain) and testing on target data (a different domain) to handle small and heterogeneous biological datasets[14]. The key insight of the algorithms described here is to fuse same-labeled data generated from different instruments or recording settings (such as voltage and frequency) to improve classification performance.

Future work will focus on precisely quantifying these relationships and developing feature-specific optimization strategies. Enhancements may include automated selection of initial transformation parameters and hyperparameter tuning to improve convergence. Extending this method to other biomedical signals, such as EEG, could validate its broader applicability. In the future, this approach could be expanded to integrate multiple datasets simultaneously, enabling comprehensive joint analyses across diverse modalities. It is proposed that this methodology for multi-source integration has broad application potential beyond ECG signals—for example, in fusing image data with physiological recordings, thereby significantly extending the utility of individual datasets. While previous ECG-based predictive models have largely concentrated on neural-network architectures[15], the use of LightGBM and XGBoost demonstrates that traditional machine-learning methods can achieve competitive performance with substantially lower computational cost. Although this study focuses on arrhythmia detection, the same framework could be applied to other ECG-detected cardiac conditions and to disorders identified via alternate cardiac signals (e.g., phonocardiogram-based fatigue detection[16]). Our enhancements in ECG-based disease identification align with recent findings that highlight the exceptional value of signal-based machine learning in critical-care settings[17]. Third, the study illustrates the promise of integrating multi-modal databases via KDE and KL divergence. By collating diverse longitudinal data from the same patient, individualized patient repositories are created. Continuously incorporating each patient's historical multi-source diagnostic records enables dynamic estimation of future disease risk, refined as

new data accrue. Aggregating these personalized repositories across a cohort also facilitates population-level analyses of specific conditions. As this framework is trained on increasingly large and heterogeneous datasets, its effectiveness for multi-source classification is expected to be strengthened. It is suggested that this strategy is particularly useful for established screening protocols that remain resource-intensive or insufficiently sensitive. Notably, this approach demands far fewer computational resources than contemporary neural-network solutions, although it is not claimed to universally outperform alternative methods. Indeed, the robust performance of the model in detecting prevalent diseases, coupled with the widespread adoption of wearable and handheld ECG devices, underscores its significant potential for large-scale screening initiatives[18].

In conclusion, our findings establish a solid foundation for divergence-based data fusion in medical AI applications. By addressing the challenges of multi-source data integration, this study contributes to the advancement of robust and reliable machine learning models for clinical decision-making. With continued development, these techniques hold the potential to improve diagnostic accuracy and facilitate personalized healthcare solutions.

Methods

This study is conducted following the workflow described in Figure 1. It is organized mainly in four steps that will be detailed in this section: data collection, data processing and feature extraction, data fusion, and machine learning-based classification.

Construction of in-house datasets

The current study first uses internal datasets comprising ECG recordings from two healthy cohorts and one arrhythmia cohort to evaluate the data fusion algorithms. Each healthy cohort includes 1,000 ECG recordings, while the arrhythmia cohort consists of 2,000 clinical ECG recordings. The recordings of a

healthy cohort were captured using a 12-lead ECG machine at a sampling rate of 500 Hz (Nihon Kohden ECG-2350, Tokyo, Japan), and the other healthy cohort and arrhythmia recordings were collected at a sampling rate of 1,000 Hz (Mindray BeneHeart R12A, Mindray Bio-Medical Electronics, Shenzhen, China). The arrhythmia cohort includes multiple forms of dysrhythmic conditions such as ectopic beats, tachycardia, bradycardia, sinus arrest, premature beats, and atrial fibrillation. Each recording is a 10-second 12-lead ECG. A total of 4,000 ECG recordings, representing 4,000 individual subjects (one recording per subject, see Supplementary Table S1), were randomly selected from a large pool of 17,363 recordings collected by the Department of Emergency Medicine at The People's Hospital of Baoan Shenzhen during 2022 and 2023. The initial diagnoses were made by the software of the ECG machine and confirmed by two experienced ECG experts. The study was approved by the Ethics Committee (reference number: BYL20240558), the institutional review board (IRB) of Shenzhen Baoan People's Hospital, and all participants signed an informed consent statement before data collection.

PTB-XL datasets and data extraction for the study

The PTB-XL database, released by the Philipps-Universität Marburg and Technische Universität Berlin, is one of the largest publicly available ECG datasets, and maintains 21,837 clearly annotated 10-second length 12-lead ECG waveform recordings from 18,885 patients and multiple different types of ECG machines[19]. In addition to the recordings from healthy subjects, the datasets include arrhythmia and various other cardiac conditions such as myocardial infarction. In the current study, we constructed four ECG datasets derived from the PTB-XL database, each of which contains 1000 recordings randomly sampled from the arrhythmia subset to match the size of the in-house datasets. Two datasets consist of recordings from healthy controls, while the other two include recordings from arrhythmia patients. Each group features different recording formats and sampling frequencies as summarized in Supplementary Table S2.

Data processing

All ECG recording data are subject to a sequential process of sampling rate normalization, down-scaling (reduction of dimensionality) coupled with resampling, denoising, and slicing of ECG signals before selection and extraction of features for analysis. Briefly, to simplify computational demands and streamline the processing framework, the sampling rates of all ECG recordings are adjusted to 500 Hz for both in-house and PTB-XL-derived datasets using resampling methods, followed by denoising by Fast Fourier Transform (FFT). In this regard, a defined threshold is used to identify and eliminate noisy elements in the data. To precisely detect the rate variability of heartbeats, an appropriate window size and sampling frequency are used to mark the regions of interest within each ECG signal to isolate heartbeat peaks. This adjustment ensures that each processed ECG signal extends uniformly to 500 samples and thus establishes a consistent baseline for further examination.

Feature selection and extraction

A total of nine specific ECG-related features (Table 5) are selected for extraction. These features cover a wide spectrum of cardiac activity and function and are extensively applied in medical signal processing to characterize ECG dynamics and detect pathological conditions[20-28]. Different feature types capture ECG signal variations across multiple scales. Each of these features has been previously validated in clinical research for their diagnostic value in various cardiac conditions. Entropy-based features (Approximate Entropy, Sample Entropy, Permutation Entropy, and Spectral Entropy) quantify the regularity and randomness of ECG signals in both time and frequency domains, making them highly sensitive to arrhythmic events and spectral energy redistribution[20,21,28]. Fractal dimensions (Higuchi, Katz, and Petrosian) capture the self-similar, multi-scale complexity of ECG waveforms, reflecting waveform fragmentation and structural abnormalities induced by pathological changes[23-25]. Detrended Fluctuation Analysis (DFA) extracts long-range correlation exponents from RR-interval fluctuations, revealing autonomic nervous system influences on heart-rate variability[27]. Singular Value Decomposition entropy

(SVD-En) evaluates the global information content of embedded ECG trajectory matrices, providing a holistic structural signature for conditions such as myocardial infarction[27]. In conclusion, this unified set of nine features is indispensable for capturing the full spectrum of ECG variations and thus enhancing diagnostic accuracy. All features are extracted from the refined datasets using the python library antropy for subsequent model training and analysis[29].

Data Fusion

The main technique used in our data fusion process is explained in this subsection. To clarify the approach, we illustrate it using an example involving data from two distinct sources, with the sets of features from these sources denoted by X and Y, respectively. The primary objective of the data fusion method is to determine an affine transformation that maps set Y to a new set $^{Y'}$ in such a way that the distributions of elements in X and $^{Y'}$ become similar. In other words, our goal is to find appropriate values for a $d \times d$ matrix C and a d-dimensional vector D, where d denotes the feature dimension, such that the following approximation is satisfied:

$$\mathrm{Law}_{y\in Y}\,(Cy+D) \approx \mathrm{Law}_{x\in X}(x) \tag{1}$$

In order to solve equation (1), We adopt a bifurcated strategy, treating Gaussian and non-Gaussian distributed features distinctly to optimize the fusion process for each case.

For Gaussian distributed features, we leverage the property that Gaussian distributions retain their natures when subjected to an affine transformation. The explicit formula for the optimal transformation parameters C^* and D^* is as follows: $= \mu_X - \sigma_X^{1/2}\sigma_Y^{-1/2}\mu_Y$

$$C^* = \sigma_X^{1/2}\sigma_Y^{-1/2}, \qquad D^* = \mu_X - \sigma_X^{1/2}\sigma_Y^{-1/2}\mu_X, \tag{2}$$

where μ_X (resp. μ_Y) and σ_X are the expectation and the covariance matrix of X (resp. Y).

When dealing with features displaying non-Gaussian distributions, we encounter two primary challenges in the context of equation (1). The first challenge lies in the difficulty of quantifying these distributions, and the second revolves around interpreting the nature of the approximation in equation (1). These two issues are addressed in the following two steps:

1. Employ KDE method to estimate the continuous distribution of elements in both sets X and Y. For instance, in the case of N data points, denoted as $x_1,\ldots,x_N$, the KDE-estimated density $\hat{f}$ is given by:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right),$$

where σ is the bandwidth.

2. Employ the KL divergence to quantify the "distance" of two estimated distributions. Given two probability density functions f and g defined in $\mathbb{R}^d$, the KL divergence from f to g is given by:

$$D_{\mathrm{KL}}(f\,||g) = \int_{\mathbb{R}^d} f(x) \log\left(\frac{f(x)}{g(x)}\right) dx.$$

Using the KDE method to compute the estimated continuous distribution of set X as $\hat{f}_X$ and to denote the estimated continuous distribution of the transformed set $CY+D$ by $\hat{f}_Y(C,D)$, We determine the optimal transformation parameters C and D in this non-Gaussian case by solving the following optimization problem associated with (1):

$$\inf_{C,D} D_{\mathrm{KL}}\left(\hat{f}_X \,||\, \hat{f}_Y(C,D)\right). \tag{3}$$

We employ the gradient descent method to solve equation (3), where the gradient of the loss function with respect to (C,D) is given by:

$$\nabla D_{\mathrm{KL}}\left(\hat{f}_X \,||\, \hat{f}_Y(C,D)\right) = -\int_{\mathbb{R}^d} \frac{\nabla \hat{f}_Y(C,D)(x)}{\hat{f}_Y(C,D)(x)} \hat{f}_X(x)\, dx. \tag{4}$$

The complete procedure for the data fusion method, applicable to both Gaussian and non-Gaussian scenarios, is summarized in algorithm 1 below. In this algorithm, the gradient is calculated using equation (4), and the stopping criterion is based on the resulting KL divergence (or gradient norm) falling below a chosen threshold.

Algorithm 1:Data Fusion Algorithm

Input: Feature sets X and Y;
if X and Y are Gaussian distributions **then**
  Return (C^*, D^*) by evaluating equation (2);
end
else
  Compute $\hat{f}_X \leftarrow \mathrm{KDE}(X)$;
  Initialize (C, D) and choose a learning rate $\tau > 0$;
  while the stopping criterion is not met **do**
    $(C, D) \leftarrow (C, D) - \tau \nabla D_{\mathrm{KL}}\left(\hat{f}_X \,\|\, \hat{f}_Y\right)(C, D)$;
  end
  Return $(C^*, D^*) = (C, D)$.
end

Classification procedure

After the feature-based fusion process, we start the classification task. The dataset of fused features is fed into the LightGBM and XGBoost models, which recognize the inherent patterns and features of the merged dataset to classify data points into predefined categories accurately. For each feature, we conduct 10-fold cross-validation (i.e., N = 10. In each fold, the same number of data points from the healthy dataset (labeled as "0") and the diseased dataset (labeled as "1") is used to ensure a balanced representation of each category. Finally, we combine the nine features for joint analysis. The effectiveness of the classification of the models is quantified by the accuracy metric, which reflects the proportion of correctly classified instances. In this way, we are able to assess the consistency of classifier performance across different scenarios, providing a comprehensive picture of their predictive power.

Data Availability

The in-house data used in this study are available from the corresponding author upon reasonable request and subject to applicable ethical and legal restrictions. Due to the sensitive nature of the data, access may be granted on a case-by-case basis following institutional and regulatory approvals.

Code Availability

The code used in this study can be available from the corresponding author on a reasonable request and subject to an institutional approval.

Acknowledgment

This work was supported by the Sanming Project of Medicine in Shenzhen (No. SZSM202206006 to Y.G. & Q.D.) and a seed grant from Aerie Intelligent Technology Corp.

Author Contributions

T.T., K.L., Y.G., and Z.T. conceived the study. T.T. and K.L. designed the model and supervised the operations. Y.G. and Q.D. collected data and participated in experiment design. J.D. contributed to data processing of both in-house and PTB-XL data. B.S. developed the code, conducted the research, and drafted manuscript. T.T., K.L., Z.Q., and Z.T. revised and finalized the manuscript. All authors reviewed the manuscript and approved its final version for publication.

Competing Interests

Z.T. is the founder of Aerie Intelligent Technology Corp (USA). The remaining authors declare no competing interests.

References

1 Arora, A. *et al.* The value of standards for health datasets in artificial intelligence-based applications. *Nat Med* **29**, 2929–2938 (2023). https://doi.org/10.1038/s41591-023-02608-w

2 Hannun, A. Y. *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* **25**, 65–69 (2019). https://doi.org/10.1038/s41591-018-0268-3

3 Sahoo, S., Dash, M., Behera, S. & Sabut, S. Machine Learning Approach to Detect Cardiac Arrhythmias in ECG Signals: A Survey. *Irbm* **41**, 185–194 (2020). https://doi.org/10.1016/j.irbm.2019.12.001

4 Chen, T. Q. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016). https://doi.org/10.1145/2939672.2939785

5 Ke, G. L. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30 (Nips 2017)* **30** (2017).

6 Schwabe, D., Becker, K., Seyferth, M., Klass, A. & Schaeffter, T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *NPJ Digit Med* **7**, 203 (2024). https://doi.org/10.1038/s41746-024-01196-4

7 Arcobelli, V. A. *et al.* FHIR-standardized data collection on the clinical rehabilitation pathway of trans-femoral amputation patients. *Sci Data* **11**, 806 (2024). https://doi.org/10.1038/s41597-024-03593-6

8 Sharma, L. D. & Sunkaria, R. K. Inferior myocardial infarction detection using stationary wavelet transform and machine learning approach. *Signal Image and Video Processing* **12**, 199–206 (2018). https://doi.org/10.1007/s11760-017-1146-z

9 Kim, J. & Scott, C. D. Robust Kernel Density Estimation. *Journal of Machine Learning Research* **13**, 2529–2565 (2012).

10 Tseng, Y.-S. & Yang, M.-H. Using Kernel Density Estimation in Knowledge Distillation to Construct the Prediction Model for Bipolar Disorder Patients. *Applied Sciences* **13** (2023). https://doi.org/10.3390/app131810280

11 Cui, J., Zong, L., Xie, J. & Tang, M. A novel multi-module integrated intrusion detection system for high-dimensional imbalanced data. *Applied Intelligence* **53**, 272–288 (2022). https://doi.org/10.1007/s10489-022-03361-2

12 Zhou, R., Gueaieb, W. & Spinello, D. A Model-Free Kullback–Leibler Divergence Filter for Anomaly Detection in Noisy Data Series. *Journal of Dynamic Systems, Measurement, and Control* **145** (2023). https://doi.org/10.1115/1.4056105

13 Farahani, A., Voghoei, S., Rasheed, K. & Arabnia, H. R. in *Advances in Data Science and Information Engineering Transactions on Computational Science and Computational Intelligence* Ch. Chapter 65, 877–894 (2021).

14 Orouji, S., Liu, M. C., Korem, T. & Peters, M. A. K. Domain adaptation in small-scale and heterogeneous biological datasets. *Sci Adv* **10**, eadp6040 (2024). https://doi.org/10.1126/sciadv.adp6040

15 Al-Zaiti, S. *et al.* Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nature Communications* **11** (2020). https://doi.org/10.1038/s41467-020-17804-2

16 Ma, X. X., Su, X. H., Ge, H. M. & Chen, Y. R. PCG-based exercise fatigue detection method using multi-scale feature fusion model. *Computer Methods in Biomechanics and Biomedical Engineering* (2024). https://doi.org/10.1080/10255842.2024.2406369

17 Lin, C. S. *et al.* AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial. *Nat Med* **30**, 1461–1470 (2024). https://doi.org/10.1038/s41591-024-02961-4

18 Al-Alusi, M. A., Ding, E., McManus, D. D. & Lubitz, S. A. Wearing Your Heart on Your Sleeve: the Future of Cardiac Rhythm Monitoring. *Curr Cardiol Rep* **21**, 158 (2019). https://doi.org/10.1007/s11886-019-1223-8

19 Wagner, P. *et al.* PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data* **7** (2020). https://doi.org/10.1038/s41597-020-0495-6

20 Bandt, C. & Pompe, B. Permutation entropy: A natural complexity measure for time series - art. no. 174102. *Physical Review Letters* **88** (2002). https://doi.org/10.1103/PhysRevLett.88.174102

21 Bruce, E. N. *Biomedical signal processing and signal modeling*. (Wiley, 2001).

22 Goldberger, A. L. *et al.* Fractal dynamics in physiology: alterations with disease and aging. *Proc Natl Acad Sci U S A* **99 Suppl 1**, 2466–2472 (2002). https://doi.org/10.1073/pnas.012579499

23 Higuchi, T. Approach to an Irregular Time-Series on the Basis of the Fractal Theory. *Physica D-Nonlinear Phenomena* **31**, 277–283 (1988). https://doi.org/Doi 10.1016/0167-2789(88)90081-4

24 Katz, M. J. Fractals and the Analysis of Waveforms. *Computers in Biology and Medicine* **18**, 145–156 (1988). https://doi.org/Doi 10.1016/0010-4825(88)90041-8

25 Petrosian, A. Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns. *Eighth Ieee Symposium on Computer-Based Medical Systems*, 212–217 (1995). https://doi.org/Doi 10.1109/Cbms.1995.465426

26 Pincus, S. M. Approximate entropy as a measure of system complexity. *Proc Natl Acad Sci U S A* **88**, 2297–2301 (1991). https://doi.org/10.1073/pnas.88.6.2297

27 Rangayyan, R. M. & Krishnan, S. *Biomedical Signal Analysis*. (John Wiley & Sons, Inc., 2015).

28 Richman, J. S. & Moorman, J. R. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology* **278**, H2039–H2049 (2000).

29 Vallat, R. *AntroPy: entropy and complexity of (EEG) time-series in Python*, <https://github.com/raphaelvallat/antropy> (2025).

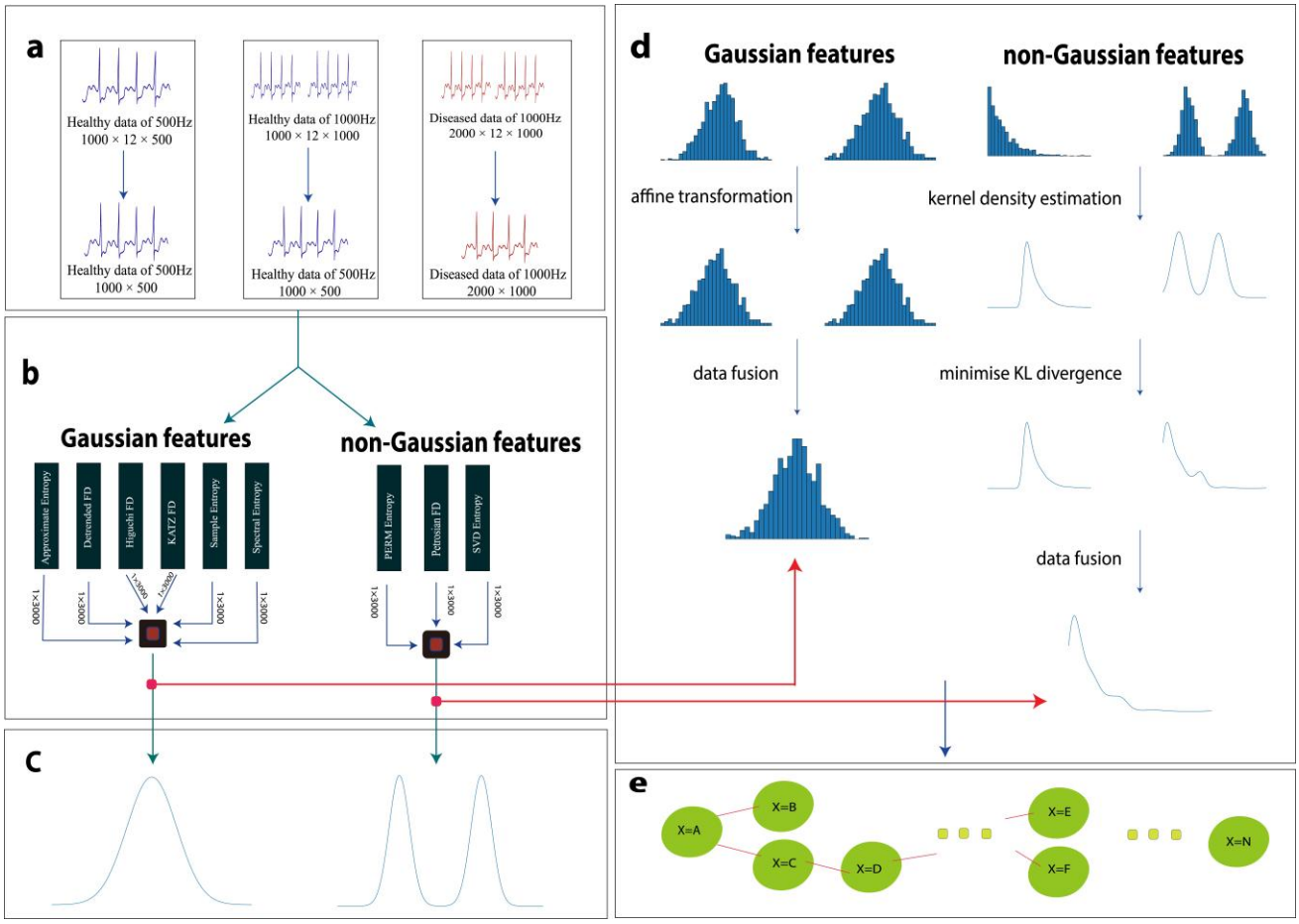


Figure 1. A schematic overview of the workflow. (**a**) Initial normalization to convert all 12-lead ECG recordings to one-dimensional signal using weighted averaging. (**b**) Extraction of six Gaussian and three non-Gaussian features. (**c**) Categorized into Gaussian and non-Gaussian features using non-parametric estimation. (**d**) Gaussian features are applied the rigid affine transformation, while non-Gaussian features are processed using KDE followed by divergence optimization. (**e**) All features processed are fed into LightGBM and XGBoost to train the model.

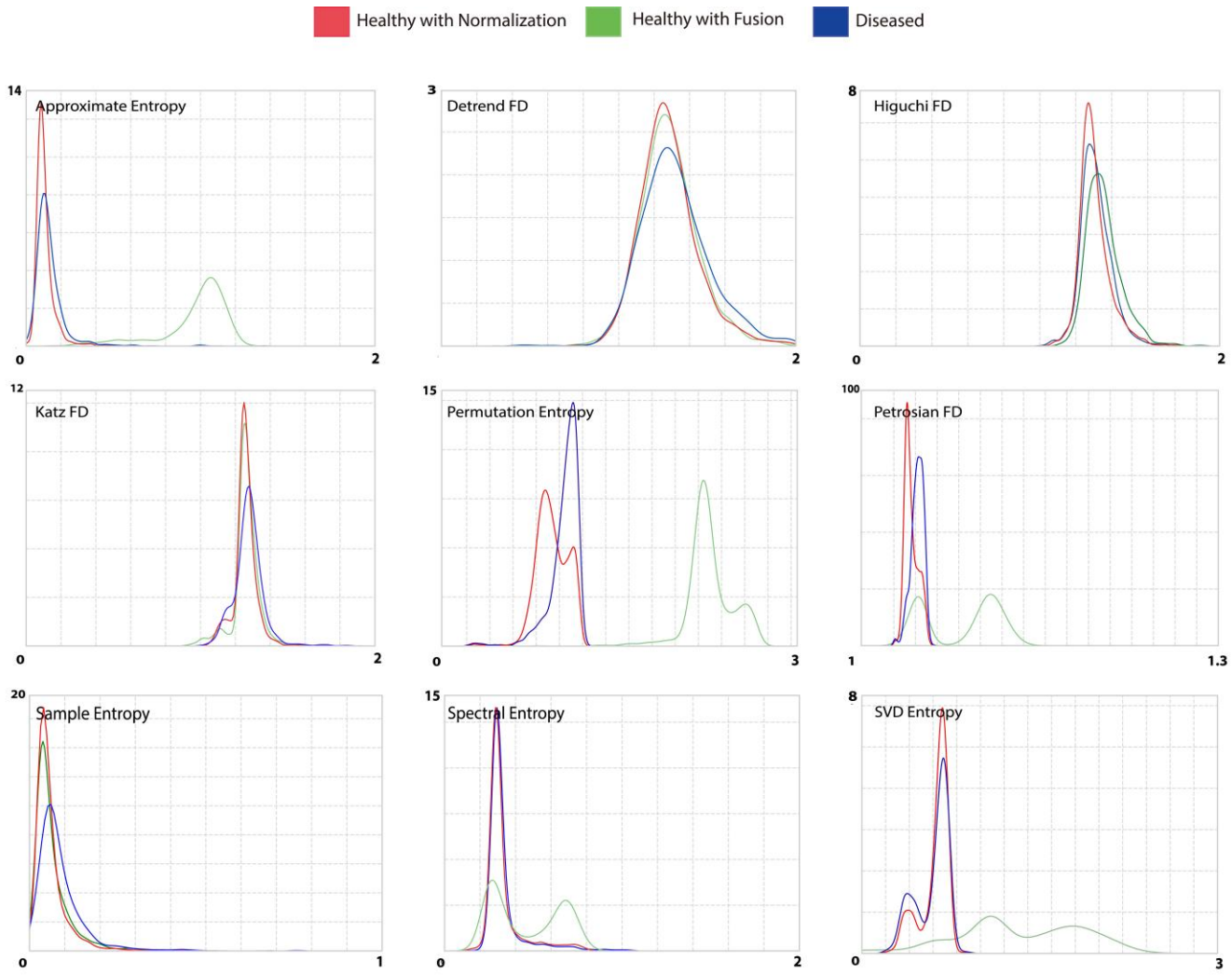


Figure 2. KDE estimated feature distributions in the in-house-derived dataset. Kernel density estimates are shown for nine signal features that include entropy measures and fractal dimensions for healthy (red, standard normalization; green, divergence-based fusion) and diseased (blue) recordings as indicated in each panel. Approximate entropy in healthy samples is tightly clustered at low values, while the diseased distribution is broader. Features such as detrended and Higuchi fractal dimensions show substantial overlap across the different methods. In contrast, non-Gaussian descriptors—Permutation entropy, Petrosian fractal dimension, and Singular Value Decomposition entropy exhibit sharper, more distinct peaks under the fusion method, with improved separation from disease distributions.

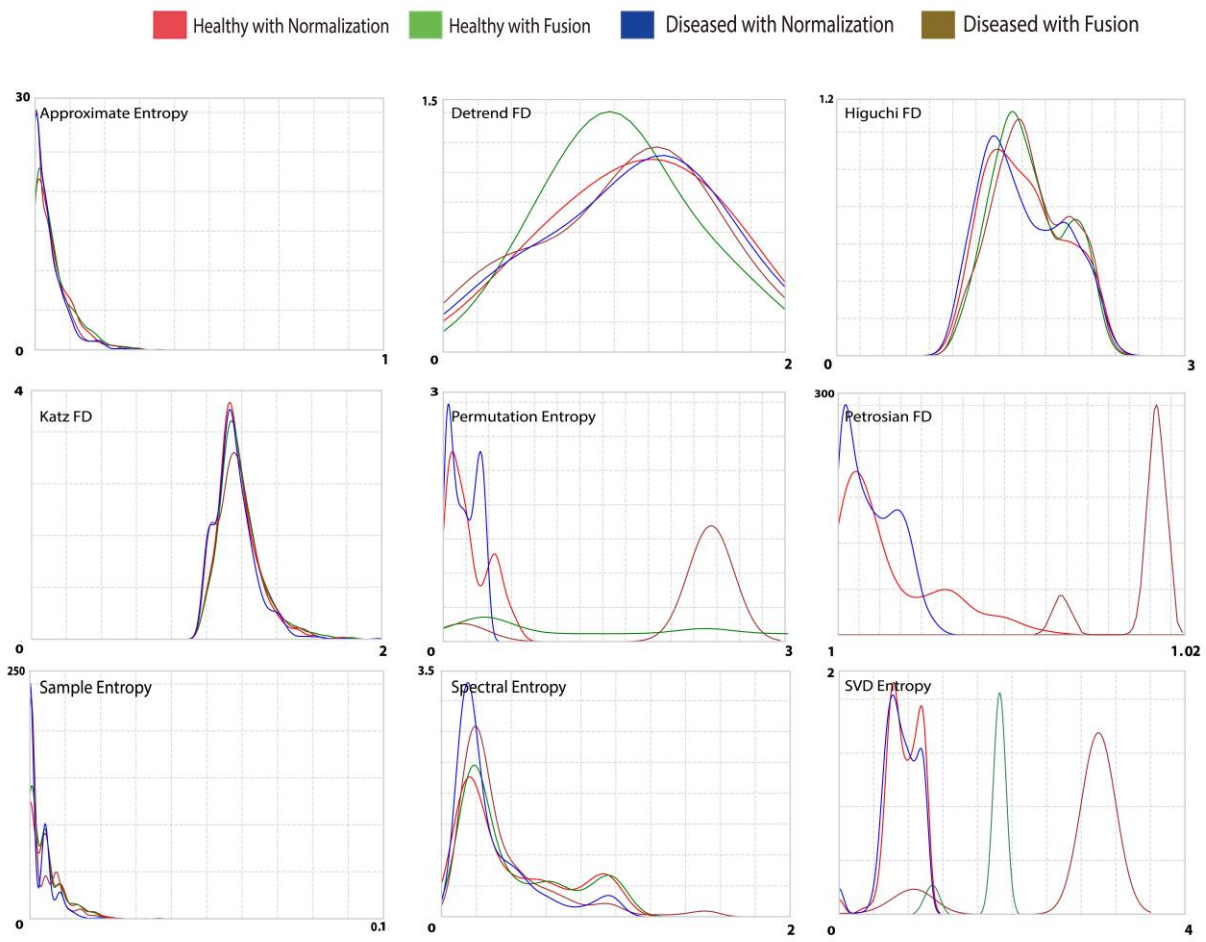


Figure 3. KDE estimated feature distributions in the PTB-XL-derived datasets. Kernel density estimation are shown for nine ECG signal features that include entropy measures and fractal dimensions—for healthy (red, standard normalization; green, divergence-based fusion) and diseased (blue, standard normalization; yellow, divergence-based fusion) recordings. Approximate entropy and sample entropy in healthy recordings are sharply concentrated at low values, while the corresponding diseased distributions display heavier tails. Detrended and Higuchi fractal dimensions and spectral entropy show notable overlap across preprocessing methods. In contrast, non-Gaussian descriptors, Permutation entropy, Petrosian fractal dimension and Singular Value Decomposition entropy exhibit more distinct and separated peaks under fusion, reducing overlap between healthy and diseased groups. Note that within the displayed intervals, the estimated density of permutation entropy (range 0–3) and

Petrosian fractal dimension (range 1–1.2) accounts for only approximately 15% and 3% of the total mass, respectively (this choice is made to enhance the compactness and clarity of the figure).

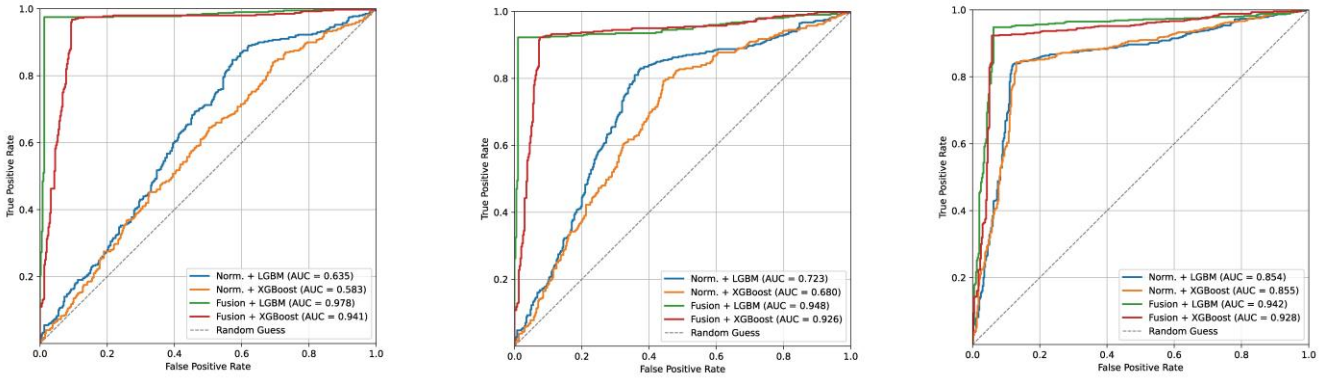


Figure 4. ROC curves for all experiments on 3 datasets. The left panel (in-house data) shows that normalization yields modest discrimination (AUC=0.635 for LightGBM and 0.583 for XGBoost) while fusion improves results (AUC=0.978 and 0.941, respectively); the middle panel (PTB-XL data) likewise demonstrates higher AUC for fusion (0.948 for LightGBM, 0.926 for XGBoost) than for normalization (0.724 and 0.680); and the right panel (combined in-house+PTB-XL) confirms that fusion (AUC=0.942 for LightGBM, 0.928 for XGBoost) consistently outperforms normalization alone (AUC=0.854 and 0.856), with the dashed diagonal in each plot marking the random-guess baseline (AUC=0.5).

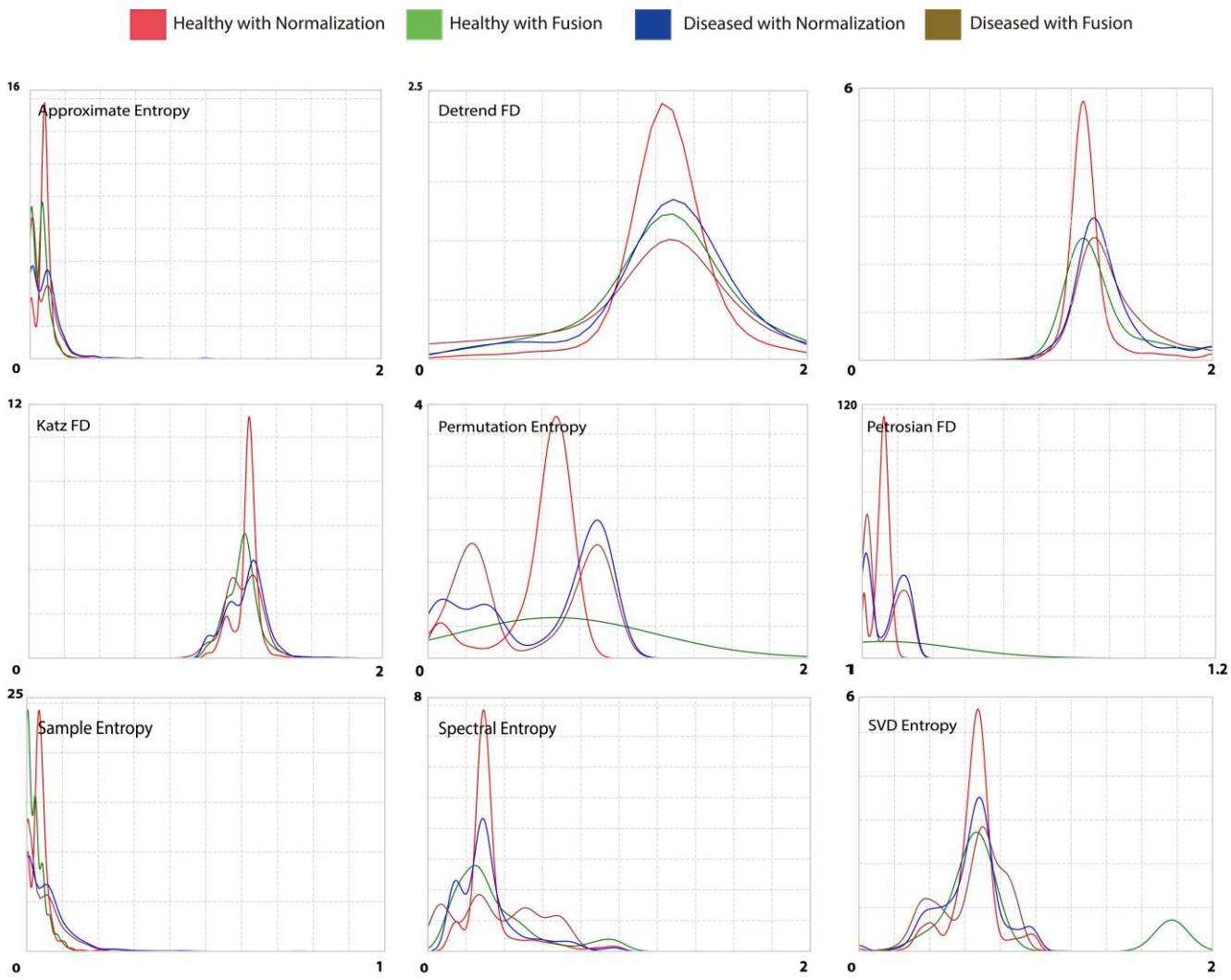


Figure 5. KDE estimated feature distributions in the aggregation-derived dataset. KDEates of nine ECG descriptors (entropy measures and fractal dimensions) for healthy (red, standard normalization; green, divergence-based fusion) and diseased (blue, standard normalization; brown, divergence-based fusion) recordings across both datasets. Insets in and zoom low-density tails, with arrows highlighting the fusion-induced shifts in approximate and Petrosian entropy. Gaussian-type metrics—detrended and Higuchi fractal dimensions, Katz fractal dimension, Sample entropy and Spectral entropy show largely overlapping curves under both preprocessings. By contrast, non-Gaussian descriptors—permutation entropy, Petrosian fractal dimension and Singular Value Decompostion entropy exhibit narrower, more distinct peaks and reduced healthy–disease overlap when processed via fusion. This demonstrates that

divergence-based fusion consistently enhances separability of complex ECG features across heterogeneous cohorts.

Table 1. Summary of classification accuracy and false positive rate (FPR) versus false negative rate (FNR) for tested features of the in-house datasets using the LightGBM (LGBM) model (data are shown as mean (95% CI)).

Distribution	Feature	Normalization Accuracy	Normalization FPR / FNR	Fusion Accuracy	Fusion FPR / FNR
Gaussian	Approximate entropy	0.347 (0.327–0.365)	0.715 / 0.615 (0.700–0.734 / 0.597–0.630)	0.992 (0.987–0.997)	0.008 / 0.003 (0.005–0.013 / 0.002–0.003)
	Detrended fractal dimension	0.331 (0.320–0.351)	0.689 / 0.658 (0.667–0.710 / 0.642–0.674)	0.562 (0.538–0.588)	0.397 / 0.482 (0.381–0.417 / 0.462–0.504)
	Higuchi fractal dimension	0.357 (0.339–0.375)	0.622 / 0.685 (0.599–0.643 / 0.666–0.709)	0.713 (0.692–0.736)	0.286 / 0.294 (0.269–0.305 / 0.280–0.312)
	Katz fractal dimension	0.351 (0.333–0.372)	0.710 / 0.597 (0.692–0.732 / 0.579–0.620)	0.676 (0.654–0.700)	0.293 / 0.360 (0.279–0.312 / 0.343–0.383)
	Sample entropy	0.358 (0.340–0.380)	0.706 / 0.597 (0.686–0.729 / 0.580–0.621)	0.758 (0.736–0.783)	0.245 / 0.141 (0.228–0.264 / 0.139–0.177)
	Spectral entropy	0.698 (0.678–0.719)	0.333 / 0.285 (0.316–0.354 / 0.270–0.303)	0.799 (0.781–0.821)	0.168 / 0.264 (0.153–0.189 / 0.248–0.287)
Non-Gaussian	Petrosian fractal dimension	0.365 (0.350–0.381)	0.615 / 0.663 (0.598–0.636 / 0.648–0.680)	0.363 (0.349–0.380)	0.643 / 0.634 (0.629–0.659 / 0.620–0.651)
	Permutation entropy	0.341 (0.329–0.358)	0.693 / 0.624 (0.669–0.718 / 0.605–0.646)	0.982 (0.971–0.998)	0.076 / 0.010 (0.059–0.095 / 0.001–0.022)
	Singular value decomposition entropy	0.321 (0.309–0.335)	0.688 / 0.682 (0.671–0.708 / 0.663–0.705)	0.862 (0.840–0.888)	0.227 / 0.071 (0.209–0.247 / 0.054–0.092)
Mixed	Combined	0.633 (0.611–0.657)	0.618 / 0.115 (0.600–0.639 / 0.098–0.134)	0.984 (0.970–0.999)	0.012 / 0.026 (0.004–0.020 / 0.015–0.040)

Table 2. Summary of classification accuracy and false positive rate (FPR) versus false negative rate (FNR) for tested features of the PTB-XL datasets using the LightGBM (LGBM) model (data are shown as mean (95% CI)).

Distribution	Feature	Normalization Accuracy	Normalization FPR / FNR	Fusion Accuracy	Fusion FPR / FNR
Gaussian	Approximate entropy	0.524 (0.509–0.539)	0.463 / 0.494 (0.446–0.490 / 0.472–0.516)	0.654 (0.638–0.673)	0.352 / 0.259 (0.334–0.373 / 0.239–0.279)
	Detrended fractal dimension	0.604 (0.589–0.624)	0.505 / 0.339 (0.487–0.527 / 0.320–0.359)	0.610 (0.594–0.628)	0.476 / 0.380 (0.457–0.499 / 0.361–0.404)
	Higuchi fractal dimension	0.528 (0.514–0.545)	0.568 / 0.429 (0.551–0.590 / 0.411–0.448)	0.536 (0.520–0.555)	0.510 / 0.446 (0.491–0.532 / 0.428–0.466)
	Katz fractal dimension	0.573 (0.557–0.590)	0.372 / 0.435 (0.363–0.395 / 0.418–0.453)	0.570 (0.554–0.586)	0.434 / 0.436 (0.416–0.457 / 0.419–0.456)
	Sample entropy	0.608 (0.592–0.624)	0.442 / 0.375 (0.424–0.464 / 0.357–0.397)	0.792 (0.775–0.811)	0.179 / 0.230 (0.162–0.199 / 0.213–0.253)
	Spectral entropy	0.571 (0.556–0.588)	0.463 / 0.404 (0.444–0.482 / 0.385–0.425)	0.661 (0.644–0.681)	0.393 / 0.314 (0.374–0.415 / 0.296–0.336)
Non-Gaussian	Petrosian fractal dimension	0.682 (0.668–0.699)	0.400 / 0.280 (0.381–0.423 / 0.262–0.302)	0.915 (0.902–0.932)	0.011 / 0.137 (0.005–0.018 / 0.123–0.157)
	Permutation entropy	0.639 (0.626–0.655)	0.347 / 0.375 (0.329–0.369 / 0.358–0.396)	0.907 (0.893–0.925)	0.063 / 0.013 (0.052–0.077 / 0.006–0.021)
	Singular value decomposition entropy	0.529 (0.514–0.547)	0.526 / 0.446 (0.507–0.550 / 0.427–0.468)	0.906 (0.891–0.922)	0.074 / 0.107 (0.062–0.087 / 0.092–0.130)
Mixed	Combined	0.756 (0.743–0.771)	0.368 / 0.179 (0.352–0.386 / 0.162–0.197)	0.958 (0.947–0.971)	0.011 / 0.077 (0.005–0.019 / 0.063–0.096)

Table 3. Summary of classification accuracy and false positive rate (FPR) versus false negative rate (FNR) for tested features of aggregated in-house and PTB-XL datasets using the LightGBM (LGBM) model (data are shown as mean (95% CI)).

Distribution	Feature	Normalization Accuracy	Normalization FPR / FNR	Fusion Accuracy	Fusion FPR / FNR
Gaussian	Approximate entropy	0.694 (0.676–0.714)	0.315 / 0.300 (0.292–0.334 / 0.289–0.317)	0.701 (0.685–0.724)	0.318 / 0.298 (0.296–0.339 / 0.278–0.316)
	Detrended fractal dimension	0.593 (0.575–0.616)	0.429 / 0.389 (0.410–0.452 / 0.367–0.414)	0.562 (0.545–0.583)	0.456 / 0.436 (0.435–0.479 / 0.421–0.462)
	Higuchi fractal dimension	0.724 (0.709–0.744)	0.310 / 0.259 (0.294–0.328 / 0.241–0.279)	0.587 (0.571–0.609)	0.490 / 0.406 (0.471–0.516 / 0.383–0.429)
	Katz fractal dimension	0.669 (0.652–0.688)	0.319 / 0.357 (0.300–0.377 / 0.337–0.377)	0.570 (0.552–0.590)	0.428 / 0.402 (0.410–0.450 / 0.384–0.424)
	Sample entropy	0.750 (0.732–0.768)	0.226 / 0.316 (0.205–0.245 / 0.296–0.339)	0.803 (0.786–0.824)	0.277 / 0.182 (0.265–0.284 / 0.157–0.197)
	Spectral entropy	0.633 (0.616–0.651)	0.379 / 0.364 (0.361–0.400 / 0.345–0.384)	0.745 (0.727–0.765)	0.393 / 0.236 (0.374–0.416 / 0.217–0.258)
Non-Gaussian	Petrosian fractal dimension	0.794 (0.779–0.813)	0.249 / 0.171 (0.231–0.268 / 0.155–0.193)	0.943 (0.931–0.958)	0.040 / 0.098 (0.016–0.046 / 0.083–0.118)
	Permutation entropy	0.782 (0.765–0.801)	0.242 / 0.200 (0.222–0.261 / 0.183–0.220)	0.987 (0.977–0.992)	0.035 / 0.018 (0.022–0.039 / 0.016–0.029)
	Singular value decomposition entropy	0.612 (0.597–0.632)	0.465 / 0.317 (0.446–0.483 / 0.298–0.338)	0.796 (0.780–0.816)	0.063 / 0.315 (0.050–0.085 / 0.298–0.336)
Mixed	Combined	0.861 (0.847–0.877)	0.118 / 0.169 (0.102–0.138 / 0.153–0.190)	0.953 (0.941–0.966)	0.063 / 0.040 (0.048–0.082 / 0.037–0.065)

Table 4. Summary of key performance metrics (Data are shown as mean (95% confidence intervals) and t-test p values).

Experiment	Precision	F-score	Sensitivity	Specificity	$p \times 10^2$ (N=10)
In-house (Normalization + LGBM)	0.588 (0.567–0.601)	0.706 (0.688–0.741)	0.885 (0.842–0.912)	0.382 (0.356–0.410)	0.205
In-house (Fusion + LGBM)	0.970 (0.953–0.982)	0.972 (0.961–0.979)	0.974 (0.962–0.990)	0.988 (0.972–0.995)	0.112
In-house (Normalization + XGBoost)	0.541 (0.519–0.556)	0.661 (0.638–0.689)	0.849 (0.812–0.874)	0.275 (0.250–0.302)	0.024
In-house (Fusion + XGBoost)	0.887 (0.873–0.906)	0.926 (0.914–0.940)	0.968 (0.954–0.981)	0.904 (0.888–0.922)	0.068
PTB-XL (Normalization + LGBM)	0.810 (0.794–0.826)	0.815 (0.802–0.834)	0.821 (0.798–0.848)	0.632 (0.606–0.658)	0.339
PTB-XL (Fusion + LGBM)	0.987 (0.975–0.994)	0.954 (0.941–0.966)	0.923 (0.901–0.944)	0.989 (0.983–0.997)	0.452
PTB-XL (Normalization + XGBoost)	0.676 (0.658–0.697)	0.734 (0.717–0.753)	0.803 (0.789–0.830)	0.530 (0.512–0.552)	0.076
PTB-XL (Fusion + XGBoost)	0.929 (0.915–0.945)	0.925 (0.912–0.941)	0.920 (0.905–0.935)	0.930 (0.915–0.946)	0.133
Aggregated (Normalization + LGBM)	0.831 (0.813–0.852)	0.831 (0.815–0.854)	0.831 (0.808–0.860)	0.882 (0.860–0.905)	0.184
Aggregated (Fusion + LGBM)	0.972 (0.958–0.986)	0.966 (0.950–0.981)	0.960 (0.945–0.974)	0.937 (0.923–0.953)	0.606
Aggregated (Normalization + XGBoost)	0.675 (0.659–0.695)	0.749 (0.733–0.774)	0.842 (0.825–0.863)	0.871 (0.854–0.890)	0.693
Aggregated (Fusion + XGBoost)	0.850 (0.832–0.871)	0.885 (0.869–0.903)	0.923 (0.907–0.940)	0.942 (0.928–0.957)	0.088

Table 5. Selected ECG Features and related descriptions

Feature Name	Feature Description
Approximate entropy	Quantifies the regularity and complexity of ECG time series; effective for arrhythmia detection.
Detrended fractal dimension	Analyzes complexity and self-similarity of ECG signals after trend removal.
Higuchi fractal dimension	Measures time-series complexity using fractal analysis.
Katz fractal dimension	Estimates fractal dimension based on signal amplitude and path length.
Petrosian fractal dimension	Calculates fractal dimensions to distinguish different complexity levels and self-similar structures.
Permutation entropy	Evaluates complexity via ordinal patterns and value ordering.
Sample entropy	Assesses irregularity and complexity; sensitive to anomalous heart rhythms.
Spectral entropy	Quantifies randomness based on power spectrum distribution across frequency bands.
Singular value decomposition entropy	Derives complexity and information content using SVD to reveal structural anomalies.