Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Multi-Class Heart Disease Detection using ECG Images via Deep CNN Feature Extraction and Ensemble Stacking

Nomula Nagarjuna Reddy[1], Lingadally Nipun[2], Md Uzair Baba[3], Nyalakanti Rishindra[4], Thoutireddy Shilpa[5]

[1-4]Student, [5]Assistant Professor, Department of Computer Science and Engineering, B V Raju Institute of Technology, Vishnupur, Narsapur, Medak, Telangana, India.

E-mail: [1]nagarjunareddynomula2@gmail.com, [2]nipun.lingadally@gmail.com, [3]uzairbvrit@gmail.com, [4]nyalakantyrishi11@gmail.com, [5]shilpathoutireddy@gmail.com

Abstract

Cardiovascular diseases (CVDs) continue to be the number one cause of mortality across the globe, illustrating the need for trustworthy and automated diagnostic methods. Electrocardiogram (ECG) analysis is a traditional method to identify cardiac abnormalities but the existing methods based on single convolutional neural networks (CNNs) or traditional machine learning (ML) classifiers suffer from overfitting, generalizing across different datasets, and addressing class imbalance, which in turn presents a barrier to developing robust systems with clinical deployment intent. This research addresses these issues by using a hybrid ensemble framework for multi-class ECG image classification. Our hybrid ensemble framework follows the approach of using transfer learning from CNNs (VGG16, VGG19, ResNet50, and InceptionV3) for deep feature extraction, applying dimensionality reduction (via Principal Components Analysis) on the reduced features, and then classifying them using a stacking ensemble of Random Forest, XGBoost, LightGBM, Multilayer Perceptron (MLP), and Support Vector Machine (SVM), with Logistic Regression serving as the meta-learner. We augmented the classes by applying the Synthetic Minority Over-sampling Technique (SMOTE) to handle imbalanced datasets. Our trials on datasets from Pakistan, Mendeley, and Bangladesh verified the effectiveness of our model, as it scored 97.6% on accuracy, 97.59% on the F1 score, and 0.9992 on the macro-AUC score, continuously performing better than both traditional ML classifiers and individual CNNs. The findings indicate that CNN-derived features combined with different ML classifiers improve the robustness of the model, its scalability, and its ability to generalize across clinical datasets. They underscore the role of the proposed model in performing disease diagnosis in real-time from an ECG and act as part of the advanced clinical decision support.

Keywords: Electrocardiogram (ECG) Classification, Deep Convolutional Neural Networks (CNNs), Stacking Ensemble Learning, Transfer Learning, Cardiovascular Disease Diagnosis, Feature Fusion.

1. Introduction

Cardiovascular diseases (CVDs) continue to be the leading cause of death around the world; according to the World Health Organization, CVDs are responsible for an estimated

DOI: https://doi.org/10.36548/jiip.2025.3.16
Received: 17.07.2025, received in revised form: 20.08.2025, accepted: 06.09.2025, published: 16.09.2025
© 2025 Inventive Research Organization. This is an open access article under the Creative Commons Attribution 4.0 International (CC BY 4.0) License

17.9 million deaths each year. ECG is one of the methods clinicians can use to evaluate the electrical activity of the heart, and even with newer techniques and strategies, it is still very much utilized and remains the quickest and most non-invasive option. ECG signals contain the rhythm and morphology of the heart resulting in the identification of conditions such as MI, arrhythmias, and structural abnormalities. Despite this, accurate interpretation and understanding of ECGs is complex, given its inter-patient variability, noise in the signal, class imbalance, and interpretation by an expert in the field.

The last few years have seen significant accomplishments in the field of ML and DL, which have made it possible for ECG-based disease classification to produce workflow and accuracy improvements. Traditionally, most proposed methods utilized either raw 1D signals processed as-is, or designed features by hand. Meanwhile, newer methods involve taking the ECG signal and converting it to a 2D image, which helps CNNs exploit the spatial aspect of the representation in order to extract features. Despite their promise, CFR ECG models are still commonly characterized by parsimony, as models are often limited to binary classification, rely on limited training datasets (thus causing overfitting), or perform poorly across datasets acquired under differing conditions. Additionally, single-model classifiers tend to perform significantly worse than a multi-classifier architecture, while classification and detection performance are generally most robust for shallow architectures. Many methods also are not built with real-time performance in mind.

While shallow ML architectures are computationally efficient, they are unable to capture the rich and complex spatial relationships contained in ECG images. CNNs circumvent some of these issues by extracting hierarchical feature representations of the image, but individually trained CNNs are dataset-dependent and can also overfit. We combine CNN-based features in an ensemble stacking approach in order to leverage both their representational power, while applying the stability of a group of diverse learners, addressing the weaknesses of both shallow learners and single models.

To resolve these limitations, recent research has turned to hybrid and ensemble learning methods. These methods utilize Transfer Learning (TL) with pre-trained CNNs for feature extraction and different classifiers for improved decision. However, there are still problems to be solved such as class imbalance, lack of diversity in datasets, failure to employ dimensionality reduction, and inefficiencies when deploying models.

As a solution to these problems, this article presents a systematic and scalable ensemble learning framework for enhancing multi-class ECG image classification. The proposed approach uses TL and employs 4 pre-trained CNNs: VGG16, VGG19, ResNet50, and InceptionV3 for deep feature extraction using PCA for dimensionality reduction and ensemble classifiers to classify the neural embeddings. The ensemble classifiers use Random Forest, XGBoost, LightGBM, MLP, and SVM, with Logistic Regression as a meta-learner. Class imbalance is handled with SMOTE.

The framework is evaluated across three publicly available, clinically annotated datasets from Bangladesh, Pakistan, and Mendeley. This multi-source validation ensures cross-population generalizability and demonstrates the practical potential of integrating such models into lightweight, real-time clinical decision support systems.

While previous ECG classification works have either implemented a single CNN method or used traditional ML classifiers, our framework incorporates the novelty of using both deep CNN feature extraction and a stacking ensemble; by using a stacking ensemble, we can satisfy the benefit of using different feature representations in a complementary manner ultimately increasing robustness and generalization across different datasets, thereby

facilitating an important contribution to the field. No such hybrid utilization has been previously explored systematically in ECG image classification, thus establishing the novelty of this approach.

2. Literature Review

The progress made in ML and DL has transformed ECG-based cardiovascular diagnoses. Various investigations into 1D signals, 2D images, or both hybrid models have been performed. However, challenges regarding generalizability, class imbalance, interpretability, and the possibility of real time use remain.

The earliest work transposed 1D ECGs to 2D binary or vectorized forms to take advantage of CNNs. These efforts within the works of Naz et al., and Ashtaiwi et al., created errors from segmentation and limitations of the dataset [1], [2], while Fatema et al. conducted hybrid model classifications using InceptionV3–ResNet50 focused on ECG images but experienced input variability issues, along with class imbalance [3]. Other work, including that of Mhamdi et al. and Aversano et al., used CNNs and introduced ensemble strategies, limited to datasets that constrained generalization [4], [5].

Other works by Kayam et al. and Sattar et al., employed temporal models of both Bi-LSTM and CNN–LSTM, albeit limited to binary classification and small dataset applications [6], [7]. Examples from Narotam et al., and Ayano et al. indicated that the works were difficult to consolidate as multimodal, nor offered interpretable evaluations [8], [9]; while Khan et al., and Nawaz et al. used ECG data that weren't images, in ensemble or traditional ML models, ultimately limiting use [10], [11].

Transfer learning has previously been utilized to mitigate data sparsity issues, such as in studies by Gajendran et al. and Sinha et al. However, the reliance on 12-lead ECGs and lack of ECG-specific fine-tuning, limits adaptability towards wearable platforms [12], [13]. Huang et al.'s large scale CNN model derived strong AUCs but was critiqued for its clinical validity due to the absence of angiographically-verified ground truth [14].

Ensemble approaches have often shown improved performance. Karthik et al. combined DBN and XGBoost; however, spatial feature depth was lacking [15]. Mishra et al.'s MATLAB pipeline produced strong accuracy, but was poorly compatible with real-time applications [16]. Yoon and Kang's stacking ensemble showed promise, but it was limited to a single-CNN approach and was also constrained by imbalance in the dataset [17]. Mahmud et al.'s model fused 1D and 2D CNNs with transfer learning, but struggled with scalability [18].

Alsekait et al.'s multimodal method shows future improved robustness, when using MRI and ECG as modality input, although their attention modelling and generalization capacity were limited [19]. Dhara et al.'s wavelet-CNN model produced strong accuracy alone, but performed poorly and inconsistently under high variability conditions [20]. Lightweight, real-time classifiers such as Mamba-RAYOLO [21] and Akter et al.'s embedded multimodal model [22] were effective, but diagnostic capabilities were limited; therefore, they were often restricted to a single-lead ECG basis while also being subject to environmental noise sensitivities. Alsayat et al. [23] proposed a deep learning ensemble for ECG classification, where the best combination of Inception, MobileNet, and NASNetLarge achieved an F1 score of 0.9651 and a balanced accuracy of 0.9640. However, their study was restricted to a single-source dataset and faced challenges of interpretability and computational efficiency, without incorporating resampling techniques or dimensionality reduction approaches. These gaps

motivate the integration of PCA and SMOTE within a hybrid ML–DL ensemble to enhance robustness and clinical applicability.

To consolidate these findings, Table 1 provides a comparative summary of recent works on ECG-based heart disease prediction, highlighting datasets, methodologies, optimization strategies, performance, and limitations. As seen in Table 1, while prior work demonstrates credible performance, it is scattered across datasets and pays little attention to the key aspects of dealing with class imbalance and dimensionality reduction. Overall, prior work has shown the expectant portion of deep learning and ensemble for ECG classification, but the critical shortfalls consist of using single datasets, a lack of dimensionality reduction, poor treatment of imbalance, and limited viability for real-world applications. These shortcomings suggest a necessary need for a unified, strong, and generalizable ensemble framework for subsequent study and developments across various clinical contexts.

In contrast to recent research studies (2023-2025) on ensemble ECG classification that typically utilized only one dataset, and limited ensemble configurations (e.g. one CNN backbone, and no balancing or resampling techniques), the hybrid approach proposed in this paper introduces a two-phase hybrid approach. First, deep features from several transfer learning models (VGG16, VGG19, ResNet50, InceptionV3) are fused to extract complementary ECG representations. Second, the fused features were modeled using a stacking ensemble of heterogeneous machine learning classifiers with Logistic Regression, as meta-learner. Unlike the work of Alsayat et al. (2025) [23] which also achieved excellent accuracy without using PCA or SMOTE, and whose experimentation was limited to one dataset, the methodological framework proposed integrates dimensionality reduction and imbalance handling across three independent datasets for testing and comparison. The proposed approach, verified in three regions (Bangladesh, Pakistan, Mendeley), demonstrates greater robustness and generalizability than published collective research approaches.

Table 1. Comparative Summary of Related Works on ECG-Based Heart Disease Prediction

Ref	Dataset Used	Model / Approach	Feature Extraction	Imbalance Optimization	Performance Acc / F1)	Limitations
Sinha et al., 2023 [13]	PTB-XL (public ECG dataset)	DASMcC (SMOTE-augmented multi-class classifier)	Handcrafted features	SMOTE	Acc ∼91%	Not image-based, limited scalability
Yoon & Kang, 2023 [17]	Chapman Univ. & Shaoxing Hospital datasets	Multi-modal stacking ensemble	CNN features	None	Acc ∼94%	Single dataset, imbalance issues
Akter et al., 2024 [22]	PTB-XL (21,799 ECGs, 18,869 patients)	Embedded real-time system (CNN+ VGG16+BiLSTM)	Raw ECG+ DL	SMOTE applied	Acc 94.07%, F1 0.94	Limited to atrial fibrillation only
Ma & Zhang, 2024 [21]	ECG images (271 patients, 1,231 images,	Deep CNN for real-time	CNN	None	Acc ∼94%	Focused on real-time, limited

	4 classes)	classificati on				dataset
Alsekait et al., 2024 [19]	HNET-DSI, HNET-DSII, HNET-DSIII	Heart-Net multimodal DL (ECG+MRI)	CNN fusion	None	AUC ~96%	Generalization limited, weak attention modeling
Mishra & Tiwari, 2024 [16]	Heart Disease (UCI Repository)	IoT-enabled 3-layer DL + meta-heuristic optimization	Deep CNN	None	Acc ~92%	MATLAB-based, not real-time deployable
Ayano et al., 2024 [9]	PTB-XL, CODE-15%, Chapman Arrhythmia	Interpretable hybrid multichannel DL	1D + 2D ECG features	None	F1 ~93%	Complex, not suitable for lightweight deployment
Kayam, 2024 [6]	MIT-BIH Arrhyth mia Database	DL-driven heart disease prediction	Raw signal	None	Acc ~90%	Binary classificati on only
Ashtaiwi et al., 2024 [2]	ECG Images Dataset of Cardiac Patients	ECG image vectorization + classification	Image vectors	None	Acc ~88%	Segmentati on issues, dataset limited
Alsayat et al., 2025 [23]	ECG Images Dataset of Cardiac Patients	Deep ensemble (Inception, MobileNet, NASNetLarge)	CNN features	None	F1 96.5% (Acc ~96.4%)	High computatio n, no PCA/SMOTE
Proposed Work	Bangladesh, Pakistan, Mendeley datasets	Hybrid Ensemble (CNN+ PCA+ SMOTE+ Stacking ML classifiers)	CNN fusion	SMOTE+ PCA	Acc 97.6%, F1 97.59%, AUC 0.9992	Strong generalizati on, real-time deployable

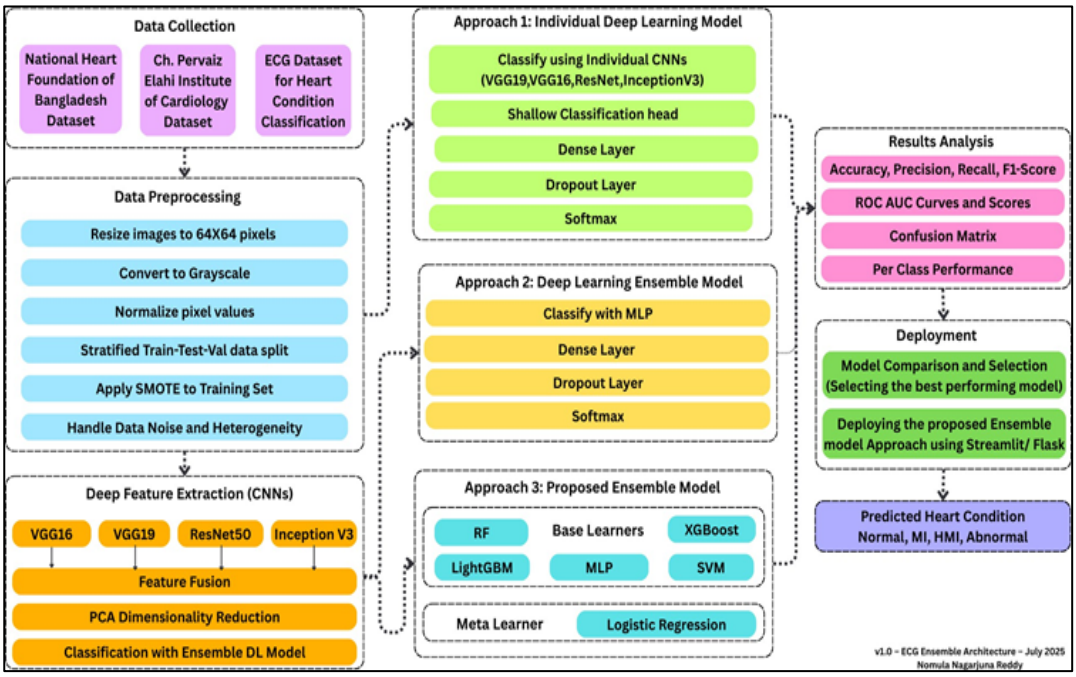


Figure 1. System Architecture

Moreover, while CNN-based models are proficient in extracting features, they are infrequently generalizable over datasets, and classical ML models are limited in reporting complex spatiotemporal ECG representations. Prior work has used PCA or SMOTE separately and typically with a single dataset, which makes them less relevant in real-world applications. However, in contrast, we used CNN-based feature extraction with PCA and SMOTE in a stacking ensemble model using three heterogeneous datasets, which addressed generalizability, class imbalance, and computational efficiency.

3. Methodology

3.1 Motivation and Proposed Solution

To address the limitations outlined above, we designed a novel hybrid ensemble learning framework that combines deep CNN based feature extraction, a stacking ensemble of diverse ML classifiers, and the LR meta-learner. In contrast to studies that study 1D signals and even a single model of CNNs [1], [4], [11], we proposed a multi-architecture feature fusion, as a method to capture richer spatial representation of ECG images. To account for class imbalance, we implemented SMOTE-based oversampling [3], [7], [17] and used PCA to reduce dimensionality and avoid overfitting [15], [20]. In contradistinction to other methods that were limited to MATLAB-only environments [16], [21], our framework is deployable on lightweight platforms enabling potential real world applications in real-time. The model is validated with three different and diverse clinical datasets to ensure generalizability, deployability, and reliability.

3.2 System Framework

The architecture of the proposed model (as shown in fig 1) follows a multi-stage pipeline. It begins with the acquisition of ECG datasets, followed by a consistent preprocessing pipeline that includes resizing images and applying SMOTE on the training data. Then the pre-

trained CNNs are selected for deep feature extraction. Once features are extracted from these models, they are concatenated and projected into PCA.

Three modeling approaches are investigated: (1) independent classification using individual CNN models with shallow classification heads; (2) deep learning ensemble modeling using an MLP classifier on the fused features; and (3) a hybrid ensemble learning model that combines diverse ML classifiers in a stacking ensemble model.

The predictions from each of these models are assessed using various evaluation metrics. The model performing best from the above assessments is used to publish the final inference model using a lightweight Streamlit (or Flask) interface for real-time inference. The expected consistent output is to predict the cardiac condition – Normal, MI, HMI, or Abnormal Heartbeat and the inferred source system will be suitable for clinical decision support system requirements.

3.3 Dataset Description

3.3.1 Dataset 1 – National Heart Foundation of Bangladesh

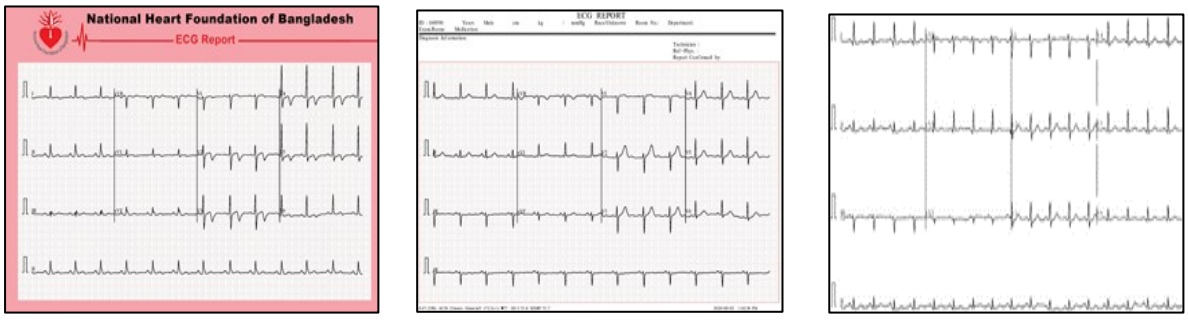
The first dataset collected from the National Heart Foundation of Bangladesh [24], began with 2898 ECG images, and was curated down to 1381 high-quality distinct images using perceptual hashing and manual curation to eliminate duplicates. All images were preprocessed at the same resolution, and the image files were labeled appropriately, following a defined folder structure. It consists of four classes, with 426 Normal, 358 Myocardial Infarction (MI), 339 Abnormal Heartbeat (AH), and 258 History of MI (HMI) images (as shown in fig2). We used this dataset to benchmark performance both on its own, and alongside other datasets of interest.

3.3.2 Dataset 2 – Chaudhry Pervaiz Elahi Institute of Cardiology, Pakistan

The second dataset [25] was comprised of 929 ECG images collected from the Chaudhry Pervaiz Elahi Institute of Cardiology in Multan, Pakistan. All images have been clinically annotated across four diagnostic categories: 284 Normal, 240 MI, 233 AH, and 172 HMI images (as shown in fig 2). Similar to the S1 dataset, this dataset includes clinically annotated images of various cardiac conditions, that represent more genuine hospital circumstances, and provide a valuable test scenario for evaluating models under resource constrained situations.

3.3.3 Dataset 3 – Mendeley ECG Dataset

The third dataset [26] consists of a total of 707 ECG images using a custom IoT-based ECG device that acquires real-time signals through electrodes placed on the chest. The dataset is well balanced over three classes: 295 Normal, 241 AH, and 171 HMI images (as shown in fig 2). The data itself highlights a real-time monitoring application and was used to evaluate model performance in wearable and IoT-based healthcare system scenarios. A notable difference was that Datasets 1 and 2 were labeled with 4 classes while Dataset 3 was labeled with only 3. To remedy this in our method of stacking ensemble while training and testing we performed class-specific probability normalization. In doing so we could still ensure the classifier learned label spaces appropriate to the dataset without deformation while maintaining some generalizability across datasets.



(a) Dataset 1 Sample (b) Dataset 2 Sample (c) Dataset 3 Sample

Figure 2. Representative ECG Images from Each Dataset

3.4 Data Preprocessing

Proper preprocessing is important for improving model performance as well as possible data consistency across different datasets. The proposed pipeline consists of steps related to traditional and deep learning-preprocessing necessary for preparing ECG images for classification.

3.4.1 Image Loading and Normalization

All ECG images were read in grayscale, and squeezed to the same dimension of 64×64 across the dataset for the preprocessing phase. We adjusted pixel values by dividing them by the maximum pixel intensity of '255', to a uniform range of [0, 1], which should help the neural networks with faster convergence rates during training and result in fewer penalties for differences in input scales. Resizing to 64x64 maximizes computational efficiency while preserving clinically-meaningful morphological patterns like QRS complexes and arrhythmias. A grayscale conversion eliminates redundancy and overlap from color channels without compromising diagnostic information, as ECG tracing is monochromatic by nature in clinical settings.

3.4.2 Train–Validation–Test Splitting

The whole dataset was split into training (70%), validation (15%), sets and testing (15%) and was stratified during the splits. Maintaining sections with class distributions was important because although the classes were set up to be balanced, there may be slight imbalances in the numbers of each class. For both types of ML and ensemble modelling, the images were flattened as image vectors after being transformed into a one-dimensional form.

3.4.3 Handling Class Imbalance with SMOTE

The training set was also re-sampled up from the imbalance that existed by using SMOTE. All images were flattened externally into 1D vectors, the minority classes were oversampled, and once the classes were resized the feature vectors were re-transformed back into image tensors for downstream processes. Both before and after the completion of SMOTE, we described the class distribution using grouped bar charts (as shown in fig 3) to visually justify the entire balancing. SMOTE was used the same way in all three datasets before training

to ensure class balance in all datasets. Not only did we receive increased recall for underrepresented classes like HMI, but we eliminated the false negatives seen in the baseline models, thus, maintaining consistency with the study's objective to reduce class imbalance.

3.4.4 Preprocessing for Deep Learning Models

As part of CNN-based feature extraction via transfer learning, the images were resized to the input size to each of the individual pre-trained models:

- VGG16, VGG19, ResNet50: 224 x 224
- InceptionV3: 299 x 299

The shielded color images were altered to RGB by duplicating the single grayscale channel for each of the three color channels. The images were also pre-processed with the frontend processing functions of each model (such as vgg16_preprocess or inception_preprocess) to batch the input images for the feature extraction portion of fine-tuning the CNN with transfer learning.

3.5 Feature Extraction and Dimensionality Reduction

3.5.1 CNN-Based Deep Feature Extraction

To obtain high-level and rich spatial features from our ECG images we employed an ensemble of four pre-trained convolutional neural network (CNN) models; VGG16, VGG19, ResNet50, and InceptionV3, all with ImageNet weights and all shown to successfully classify medical images, which will be further demonstrated in the experimental section of this study. All models were chosen because they each have complementary architectures based on medical image classification.

For each model, we used the final convolutional layers, which included global average pooling to extract a fixed-length feature vector. We omitted the head (i.e., fully connected layers) of each model by specifying include top=false, allowing only spatial features to be extracted. Each CNN model has limitations, such as VGG's depth allowing for fewer features, ResNet capturing edge patterns with its residual learning, and Inception capturing multi-scale features. However, their combination will alleviate these issues because the ensemble will preserve their complementary representations producing more robust and generalization across varying ECG patterns.

3.5.2 Feature Fusion and Ensemble Representation

The fused feature vector for each image contains varied representations gleaned from different CNNs architectures and thus enriching the model with greater generalization capabilities over datasets. These features were computed respectively for the training, validation and testing sets. The ensemble method was crucial in utilizing the complementarity of each base model while significantly improving classification performance compared to just one of the individual models

3.5.3 Dimensionality Reduction using PCA

The combined feature vectors represented high-dimensional feature spaces, and thus required caution against computational expense and overfitting. To control for these concerns, we performed a principal component analysis (PCA), also called eigendecomposition, to diminish the dimensionality while keeping most of the variance in the feature space. PCA was fitted on the training set and then applied to the validation and test sets. We also generated 2D visualizations using PCA projections to assess sample distribution (as shown in fig 3) and to visualize sample class separability before and after the application of SMOTE.

To assess a suitable dimensionality cut-off, we analyzed PCA components that ranged between 100 and 250. The model performance remained stable beyond ∼ 140 components as accuracy and AUC differed by little (<0.2%). At 150 components, we retained >97% of the variance; thus, we balanced redundancy reduction with enough discriminative information. While we could have used more components (200 - 250), no measurable benefit was seen with downstream performance. The only likely effect of having more components would be to increase the training and inference time. On this basis, we regarded 150 components as the most economical and sound cut-off point. The reduced feature representations were used as inputs for both the deep ensemble classifier and the traditional ML-based stacking ensemble (described in Section 3.4) classifier.

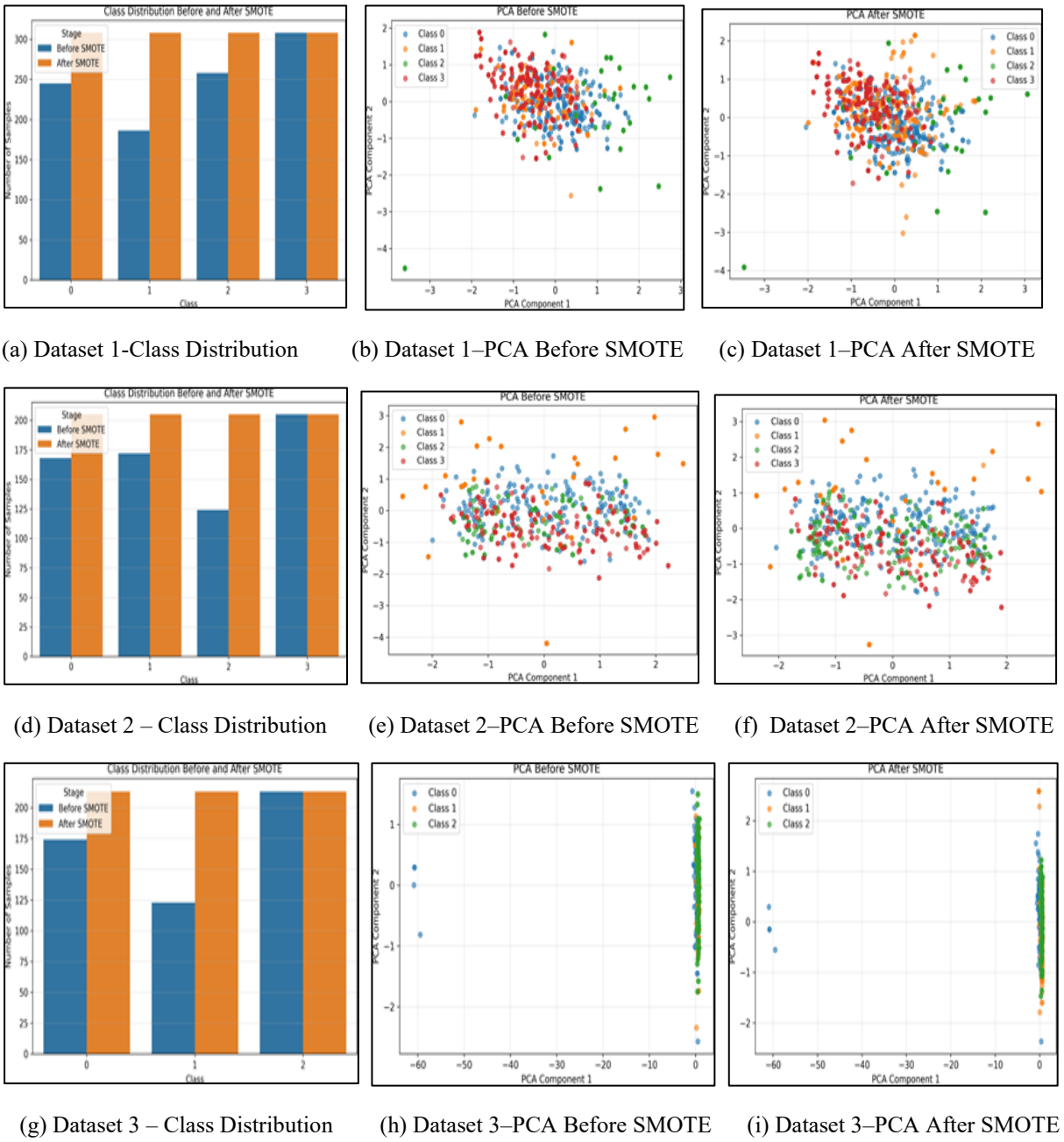
3.6 Classification Framework

This section describes the approaches used for multi-class ECG image classification. We structure the approaches into three principal pipelines: (i) each for individual deep learning models for a baseline measure of performance, (ii) a deep feature fusion pipeline to improve learning, and (iii) a stacked set of classic machine learning models using reduced deep features. Each pipeline is assessed to understand its contribution towards developing reliable detection of cardiovascular disease.

3.6.1 Individual Deep Learning Models

In order to establish a baseline and evaluate the networks' discrimination capabilities, several common pretrained models with reasonable popularity were used in isolation: VGG16, VGG19, ResNet50, InceptionV3, and EfficientNetB0. These models were selected because of their various architectural differences different depths, receptive fields and convolutional styles impacting their ability to target local and global ECG features.

For all networks, the models were adapted to the grayscale ECG images. This included resizing the images and adjusting the single channel to the RGB inputs to interact with samples. Features were extracted from the penultimate layer after global average pooling with a shallow classification head. This shallow classification head contained only dense layers and dropout layers; having shallow classifiers helps retain a computationally efficient process and maintain appropriately learned representations from their pretrained model on ImageNet.



(a) Dataset 1-Class Distribution (b) Dataset 1–PCA Before SMOTE (c) Dataset 1–PCA After SMOTE

(d) Dataset 2 – Class Distribution (e) Dataset 2–PCA Before SMOTE (f) Dataset 2–PCA After SMOTE

(g) Dataset 3 – Class Distribution (h) Dataset 3–PCA Before SMOTE (i) Dataset 3–PCA After SMOTE

Figure 3. Comparison of Class Distribution and PCA Transformation Before and After SMOTE Across Three Datasets

3.6.2 Deep Feature Fusion with CNN Ensemble

To mitigate the architectural limitations of single CNNs, and leverage their complementary feature representations, deep feature fusion was considered. The features obtained from VGG16, VGG19, ResNet50, and InceptionV3 could then be concatenated to create a single high-dimensional representation for every ECG image. This was done by exploiting the smaller fine-grained local patterns and larger semantic representations or abstractions of the ECG image attributes that could not be fully encoded in one architecture individually. Once extracted, the features were pooled to aggregate all features using average pooling and then each image was concatenated along the feature dimension.

An ensemble DL classifier was trained with the PCA reduced features in the context of an MLP architecture. The model used two fully connected layers with dropout regularization, and a softmax classifier on top of the preceding layers. The model improved robustness and generalization compared to the individual CNNs since it used the complementary strengths of multiple architectures.

3.6.3 Stacked Machine Learning Ensemble

To further improve classification performance, we implemented a stacked ensemble [27] of machine learning classifiers on the PCA-constructed deep features. The base learners were: Random Forest [28], XGBoost [29], LightGBM, Multilayer Perceptron (MLP) [30] and Support Vector Machine (SVM) [31] – These classifiers used represented different model paradigms: bagging, boosting, kernel methods, and neural computation. Their very differences were why their combination was intended to address variance, bias, and overfitting in different data subsets.

A Logistic Regression model [32][33] was utilized as the meta-learner to synthesize the predicted outcomes from the base classifiers. This was done to guarantee some interpretability at the final decision layer, and produce temporal stability in the outputs from the ensemble learner. The train-test method for appraising performance employs K-fold cross-validation and stratified 5-fold cross-validation to maintain integrity by avoiding data leakage in the training of the meta-model.

The ensemble learner [27] resulted in improvements in classifier performance, which highlighted the enhanced performance with respect to class imbalance and more complex relationships, due to the diversity and complementarity among each model.

The stacking classifier outperformed both individual CNNs, and the deep learning ensemble, producing the highest F1 score (97.59%), accuracy (97.60%), and macro AUC (0.9992) found on the curated dataset from Bangladesh. The results highlight the utility of hybrid learning that combines the best of deep neural representation with the interpretability and robustness of traditional ML.

3.6.4 Training Configuration and Implementation Details

For the DL ensemble model, we trained a feedforward network with two dense layers consisting of 512 and 256 units, using ReLU activation and dropout (0.5 and 0.3), using concatenated CNN features. The training utilized Adam, sparse categorical cross-entropy loss, a batch size of 32, and early stopping with a patience of 7, with a maximum of 100 epochs and a scheduled learning rate decay.

We ultimately chose Adam because of its adaptive learning rate and strong empirical stability with some ECG classification tasks. Initial testing with RMSProp and stochastic gradient descent (SGD) with momentum produced slower convergence and had less empirical stability with validation accuracy (fluctuations were approximately ±2-3%). AdamW performed similarly to Adam but added latency due to the overhead of updating the weights with decay. Thus, Adam was retained because it provided the best trade-off between accuracy and empirical stability with our lightweight implementation.

The feature extractor part of the pre-trained CNN was not fine-tuned or trained; it was only used as a frozen feature extractor. Images were pre-processed, and features were extracted using global average pooling.

The classical ensemble ML and DL models were trained by fitting a stacking classifier using the PCA-reduced CNN features and SMOTE balanced data, evaluated using 5-fold stratified cross validation. Also, we used the individual CNN features to train the shallow neural networks with similar training parameters but a ceiling of only 60 epochs and early stopping (patience = 5), using the shallow networks for comparative purposes as well as with the DL ensemble.

All experiments were conducted on a Windows 11 machine with an 11th Gen Intel Core i5-1135G7 CPU (2.40GHz 4 cores 8 threads), with 8 GB RAM and no GPU. The entire pipeline took approximately 2.5 to 3 hours across datasets.

Following preprocessing (resizing and converting to grayscale), each image took less than 2 ms. The frozen CNN feature extraction across the four backbone models took around 80-100 ms per image on CPU. PCA reduced the embeddings by nearly 70% (down to 150 components), greatly reducing memory footprint and speeding up any downstream training or inference. The ensemble classification took only 5-7 ms per sample resulting in an end-to-end inference time of 100-120 ms per ECG image on CPU. The total memory footprint was below 100 MB which supports that the framework is operating in near real-time value even in the absence of GPU acceleration.

3.7 Evaluation Metrics

To conduct a thorough evaluation of the classification performance of the proposed models, we will use a variety of established evaluation metrics commonly used in multi-class medical image classification, including Accuracy, Precision, Recall, F1 Score and Macro AUC.

- Accuracy measures the proportion of correctly classified samples out of all predictions.
- Precision measures the model's ability to identify positive class instances correctly, while not misclassifying any negative samples as positives.
- Recall reflects how well the model recognized all relevant instances for each class.
- F1 Score computes the harmonic mean of Precision and Recall, which can be especially useful in imbalanced datasets.
- Macro AUC (Area Under the ROC Curve) describes the model's ability to discriminate multiple classes, by determining the average AUC for all one-vs-rest comparisons.

The overall distribution ensures a balanced evaluation, and as stated above, Macro AUC is useful in depicting the model's capacity to distinguish between a variety of subtle cardiac anomalies and also provides grounding for comparison when looking at imbalanced classes overall.

The metrics applied to the proposed stacking ensemble and baseline models were done uniformly across all three datasets, enabling a like-for-like comparison with previous ECG classification studies. In addition, normalized confusion matrices (as shown in fig 6), provide comprehensive information on the per-class sensitivity, specificity and per-class error trends. Collectively these measures are exhaustive with respect to clinical reliability thus, requiring no

supplementary statistics (e.g., Cohen's Kappa or MCC) which can be inferred from the provided results.

4. Results

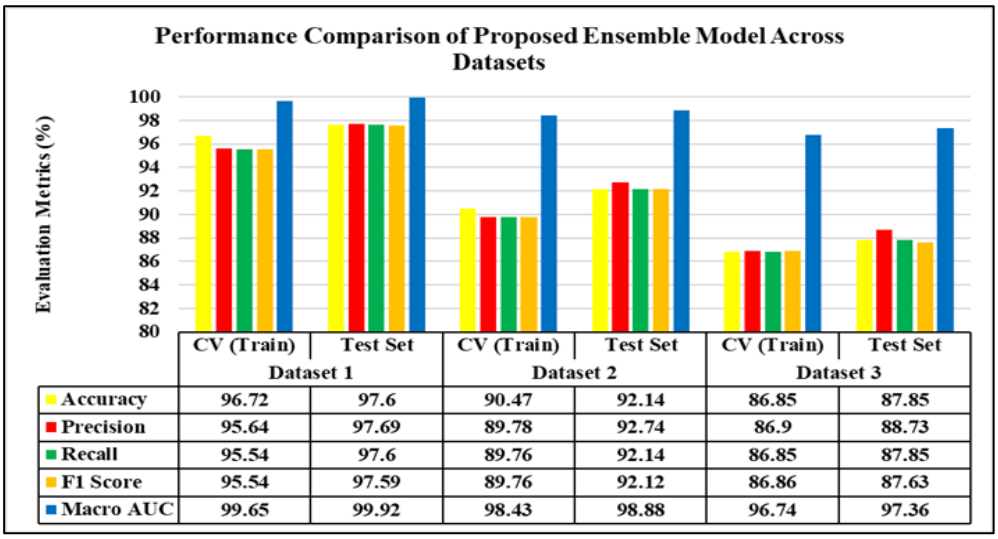
In this section, we present an extensive examination of the experimental results obtained by assessing individual deep learning (DL) models, classical machine learning (ML) classifiers, and our hybrid ensemble framework on three publicly available ECG datasets. The evaluation was performed based on four metrics: Accuracy, Precision, Recall, F1 Score, and Macro AUC.

4.1 Performance of the Proposed Hybrid Ensemble Framework

The proposed model consistently outperformed all other models across the datasets. The performance of the proposed model compared to previous models is shown in fig 4:

- **Dataset 1:** Test Accuracy 97.60%; F1 score 97.59%. Macro AUC of 0.9992.
- **Dataset 2:** Test Accuracy 92.14%; F1 score 92.12% Macro AUC of 0.9888.
- **Dataset 3:** Test Accuracy 87.85%; F1 score 87.63%. Macro AUC of 0.9736.

In order to evaluate overfitting, we compared the scores of the performance metrics between cross-validation (training) and held-out test sets across the three datasets as shown in fig 4. Overall, there is very little difference between training and test scores for all major metrics assessed. This consistency shows that the proposed ensemble can generalize well and has avoided substantial overfitting.



	CV (Train)	Test Set	CV (Train)	Test Set	CV (Train)	Test Set
	Dataset 1		Dataset 2		Dataset 3	
Accuracy	96.72	97.6	90.47	92.14	86.85	87.85
Precision	95.64	97.69	89.78	92.74	86.9	88.73
Recall	95.54	97.6	89.76	92.14	86.85	87.85
F1 Score	95.54	97.59	89.76	92.12	86.86	87.63
Macro AUC	99.65	99.92	98.43	98.88	96.74	97.36

Figure 4. Training vs Testing performance of the Ensemble Across Datasets, showing Stable Generalization with Minimal Overfitting

To assess cross-dataset generalization, the ensemble was trained separately and evaluated on the Bangladesh, Pakistan, and Mendeley datasets. The consistency demonstrated shows robustness to changes in acquisition conditions and label quality.

In addition to validating statistical reliability through 5-fold cross-validation repeated three times, the standard deviation of accuracy across folds was below 0.7% indicating stable

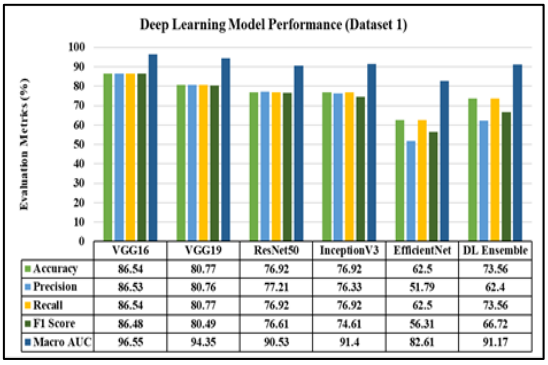
generalization. A paired t-test between the ensemble and the top performing single CNN (VGG16) indicated there was a statistically significant difference in accuracy ($p < 0.01$). The results show that the gains of the ensemble are not simply due to random variance but appear to be stable and replicable.

4.2 Individual Machine Learning Model Results

In sum, displayed in figure 5 (b, d, f) are the results for seven individual ML classifiers. Key points:

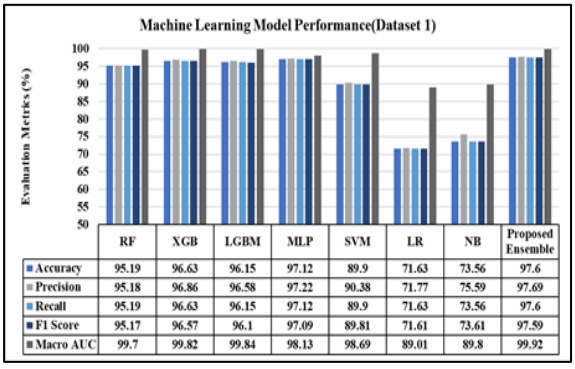
- **For dataset 1:** MLP was the top performer with an F1 score of 97.09%, with XGBoost and LGBM also performing well.
- **For dataset 2:** XGBoost had the highest individual F1 score of 89.08% with MLP (88.42%) and LGBM (88.46%) nearly in the same ballpark.
- **For dataset 3:** RF and MLP were the highest performers with F1 scores of 88.56% and 87.63%, respectively.

In summary, while the individual models all had strong results, none were able to consistently replicate the model performance achieved with the ensemble over all datasets. The ensemble solution indicated higher predictive power and robustness of the model accuracies, suggesting that combining ML classifiers from varying groups yields stronger results than individuals alone.



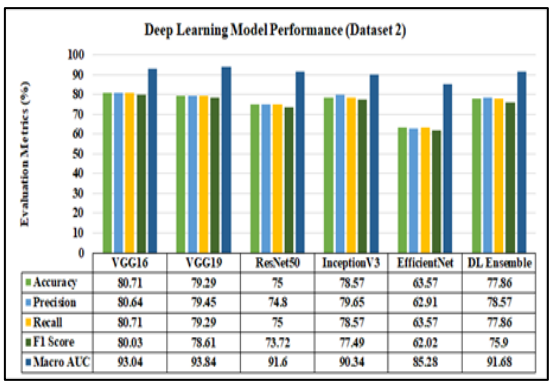
	VGG16	VGG19	ResNet50	InceptionV3	EfficientNet	DL Ensemble
Accuracy	86.54	80.77	76.92	76.92	62.5	73.56
Precision	86.53	80.76	77.21	76.33	51.79	62.4
Recall	86.54	80.77	76.92	76.92	62.5	73.56
F1 Score	86.48	80.49	76.61	74.61	56.31	66.72
Macro AUC	96.55	94.35	90.53	91.4	82.61	91.17

(a) DL Performance – Dataset 1



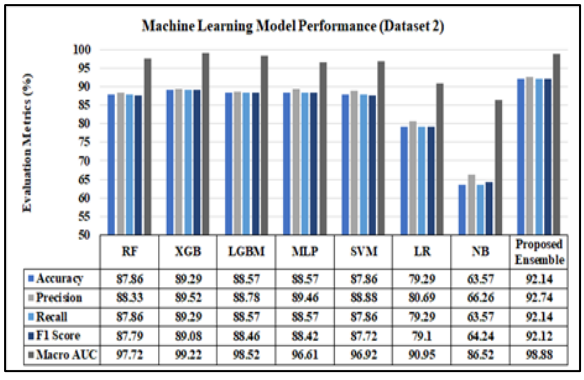
	RF	XGB	LGBM	MLP	SVM	LR	NB	Proposed Ensemble
Accuracy	95.19	96.63	96.15	97.12	89.9	71.63	73.56	97.6
Precision	95.18	96.86	96.58	97.22	90.38	71.77	75.59	97.69
Recall	95.19	96.63	96.15	97.12	89.9	71.63	73.56	97.6
F1 Score	95.17	96.57	96.1	97.09	89.81	71.61	73.61	97.59
Macro AUC	99.7	99.82	99.84	98.13	98.69	89.01	89.8	99.92

(b) ML Performance – Dataset 1



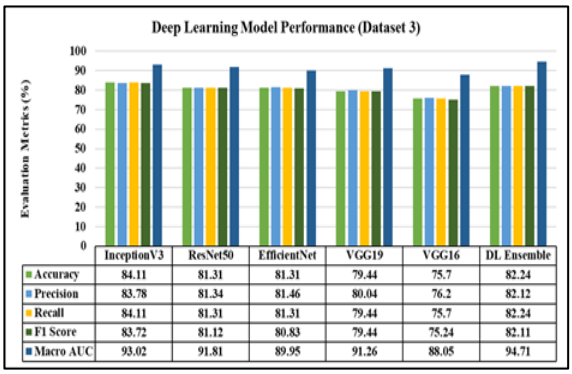
	VGG16	VGG19	ResNet50	InceptionV3	EfficientNet	DL Ensemble
Accuracy	80.71	79.29	75	78.57	63.57	77.86
Precision	80.64	79.45	74.8	79.65	62.91	78.57
Recall	80.71	79.29	75	78.57	63.57	77.86
F1 Score	80.03	78.61	73.72	77.49	62.02	75.9
Macro AUC	93.04	93.84	91.6	90.34	85.28	91.68

(c) DL Performance – Dataset 2



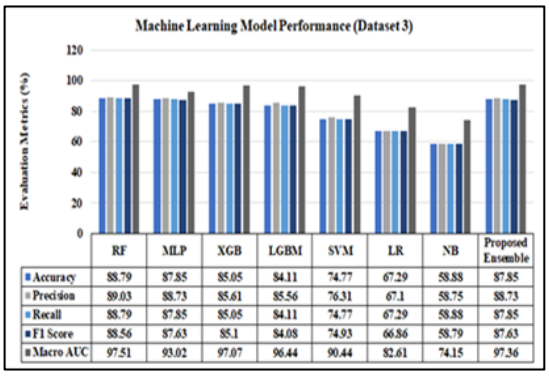
	RF	XGB	LGBM	MLP	SVM	LR	NB	Proposed Ensemble
Accuracy	87.86	89.29	88.57	88.57	87.86	79.29	63.57	92.14
Precision	88.33	89.52	88.78	89.46	88.88	80.69	66.26	92.74
Recall	87.86	89.29	88.57	88.57	87.86	79.29	63.57	92.14
F1 Score	87.79	89.08	88.46	88.42	87.72	79.1	64.24	92.12
Macro AUC	97.72	99.22	98.52	96.61	96.92	90.95	86.52	98.88

(d) ML Performance – Dataset 2



	InceptionV3	ResNet50	EfficientNet	VGG19	VGG16	DL Ensemble
Accuracy	84.11	81.31	81.31	79.44	75.7	82.24
Precision	83.78	81.34	81.46	80.04	76.2	82.12
Recall	84.11	81.31	81.31	79.44	75.7	82.24
F1 Score	83.72	81.12	80.83	79.44	75.24	82.11
Macro AUC	93.02	91.81	89.95	91.26	88.05	94.71

(e) DL Performance – Dataset 3



	RF	MLP	XGB	LGBM	SVM	LR	NB	Proposed Ensemble
Accuracy	88.79	87.85	85.05	84.11	74.77	67.29	58.88	87.85
Precision	89.03	88.73	85.61	85.56	76.31	67.1	58.75	88.73
Recall	88.79	87.85	85.05	84.11	74.77	67.29	58.88	87.85
F1 Score	88.56	87.63	85.1	84.08	74.93	66.86	58.79	87.63
Macro AUC	97.51	93.02	97.07	96.44	90.44	82.61	74.15	97.36

(f) ML Performance – Dataset 3

Figure 5. Comparison of DL and ML Models Across Datasets, Where the Ensemble Consistently Outperforms Individual Models

4.3 Individual Deep Learning Model Results

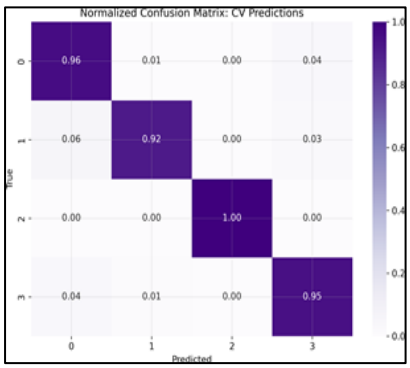
Figure 5 (a, c, e) shows the classification predictions for the individual DL models using transfer learning. On Dataset 1:

- VGG16 performed the best of all DL models, with an 86.48% F1 score and a 96.55% AUC.
- Models like InceptionV3 and EfficientNet failed to reach an F1 score of 75%, and were poor performers on Dataset 2 and Dataset 3.

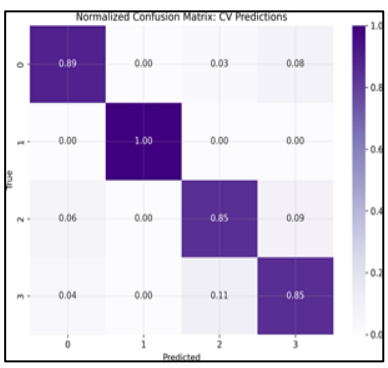
The DL ensemble was able to achieve slight improvements, but, again, the ML ensemble still performed best. For example, on Dataset 3, DL ensemble models achieved an 82.11% F1 score and a 94.71% AUC.

4.4 Confusion Matrix Analysis

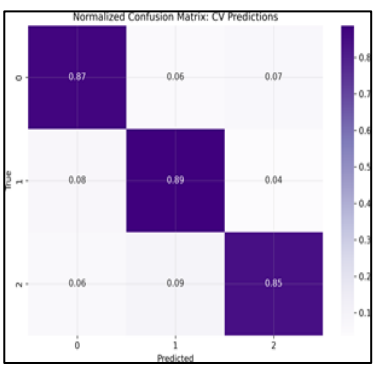
The normalized confusion matrices for the training and testing phases over all three datasets are shown in fig 6. The ensemble model demonstrates excellent class separation, especially in Dataset 1, where the model exhibits predictions with near perfection across all four classes. In the more challenging Dataset 3, the model shows over 85% diagonal percentage which reflects both strong class discrimination and reliable classification of the minority classes, such as HMI.



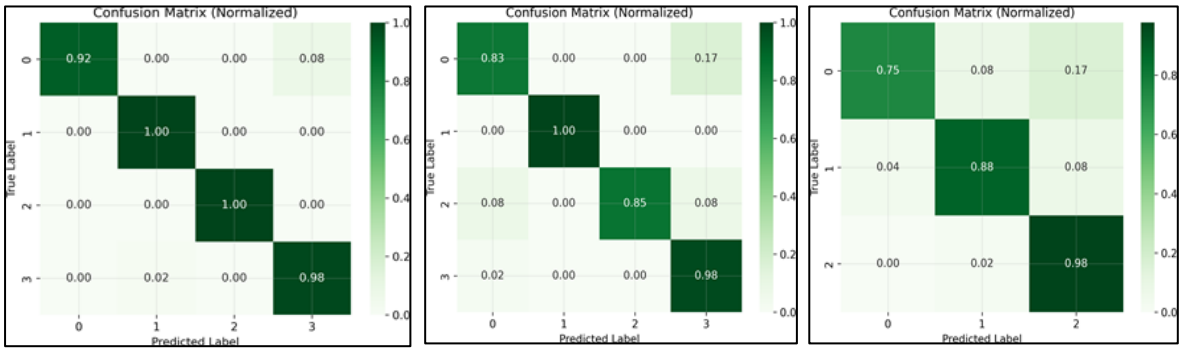
(a) Train – Dataset 1



(b) Train – Dataset 2



(c) Train – Dataset 3



(d) Test – Dataset 1 (e) Test – Dataset 2 (f) Test – Dataset 3

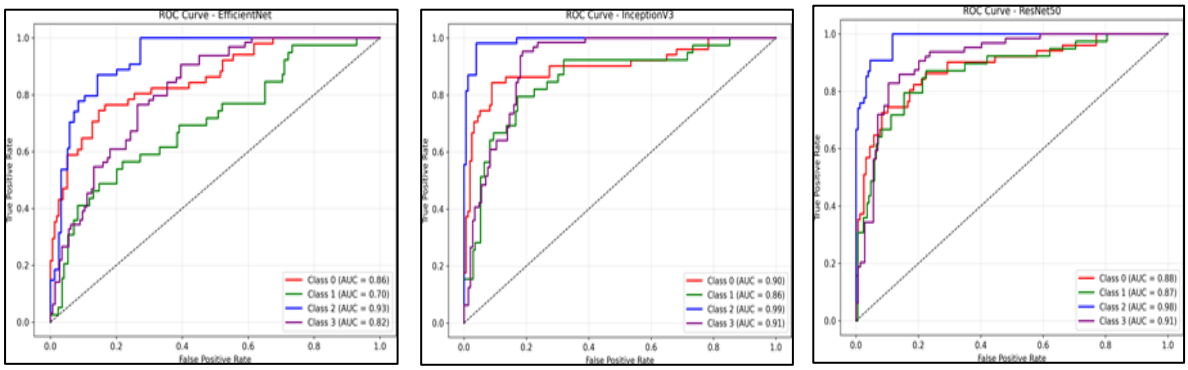
Figure 6. Normalized Confusion Matrices Showing High Class-Wise Accuracy with Few Misclassifications Across Datasets

4.5 ROC Curve Analysis

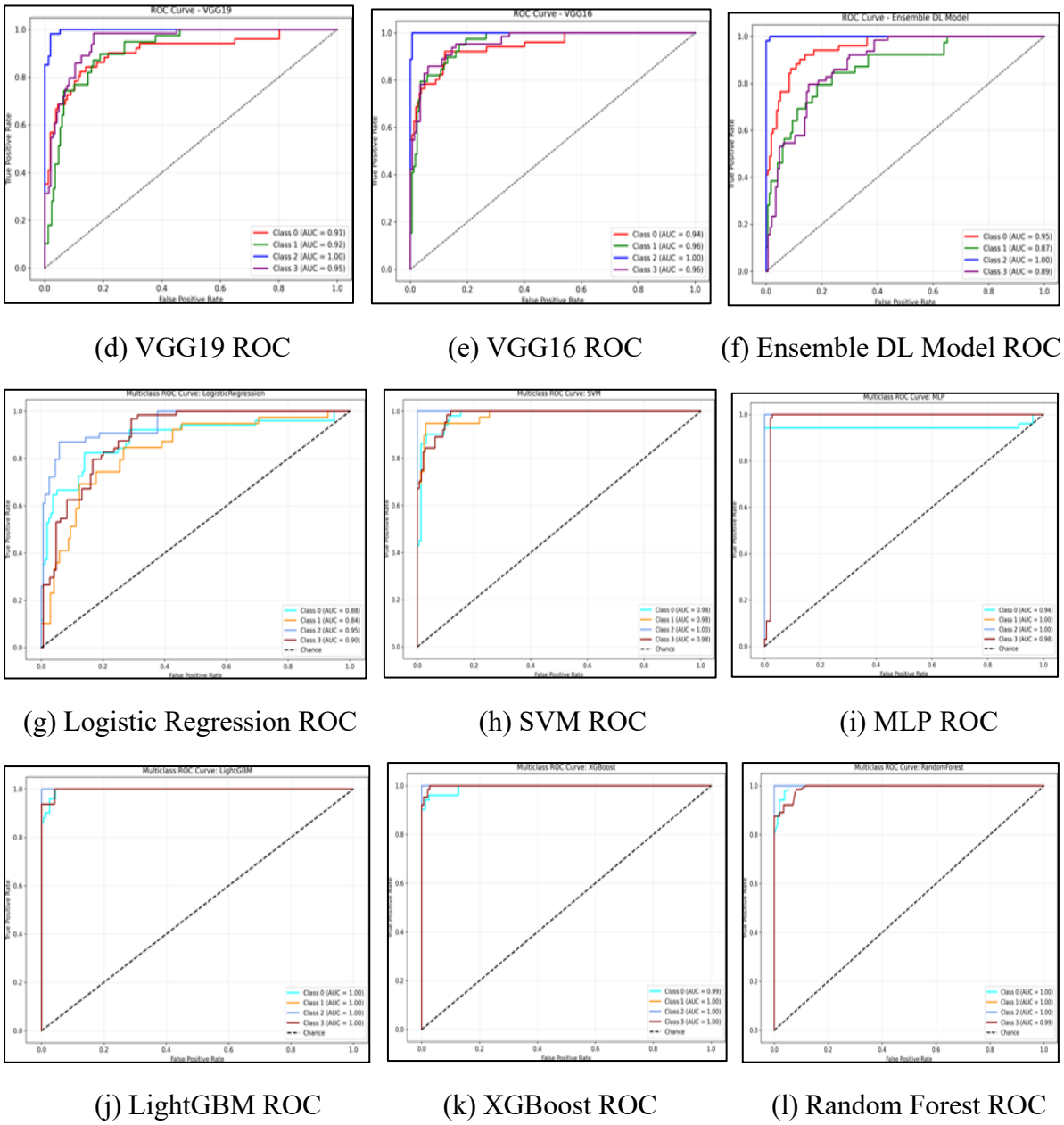
Figures 7, 8 and 9 show the 12 models' Receiver Operating Characteristic (ROC) curves on Datasets 1, 2, and 3, respectively. The proposed ensemble again provided better area under the curve (AUC) values for each of the classes consistently and had only weakly discriminative class prediction with the individual models (SVM, LR, and NB). Although XGBoost, MLP, and VGG16 produced ROC plot performance levels that were comparable to the ensemble in certain class-wise ROC plot results, they did not provide performance levels that were consistent across datasets.

5. Discussion

The experimental results from our three real-life ECG datasets validated the efficiency of the proposed hybrid ensemble learning framework. In this section, we will describe key takeaways, performance trends, and implications.



(a) EfficientNet ROC (b) InceptionV3 ROC (c) ResNet50 ROC



(d) VGG19 ROC (e) VGG16 ROC (f) Ensemble DL Model ROC

(g) Logistic Regression ROC (h) SVM ROC (i) MLP ROC

(j) LightGBM ROC (k) XGBoost ROC (l) Random Forest ROC

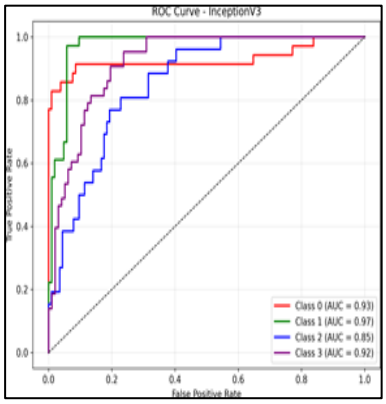
Figure 7. ROC Curves of 12 Models on Dataset 1, Where the Ensemble Achieves the Highest AUC

5.1 Superiority of the Proposed Ensemble Framework

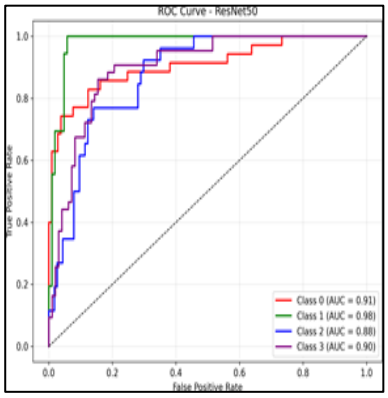
The ensemble model which takes the CNN-based features and meta-learners over diverse ML classifiers was superior across all metrics to both DL techniques and individual ML models. For instance, in Dataset 1 the test F1 score was 97.59% which is both high in discriminatory power and robustness. Even for Dataset 3 with its limited number of samples and ambiguous signals, the ensemble model exhibited an F1 score of 87.63% and an AUC score of 0.9736, which speaks to its generalizability to other distributions of data. We believe its superior performance is attributable to:

- The CNN-based features are inherently richer in patterns being spatial in their extraction of the ECG signals

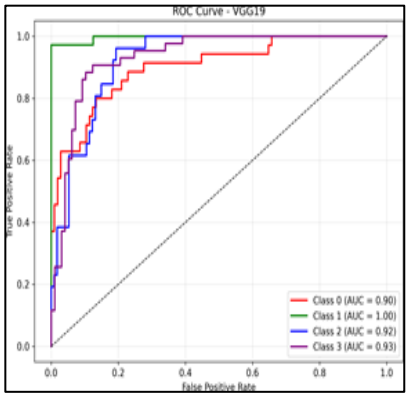
- PCA dimensionality reduction was achieved to minimize overfitting and clean the classification signal dimension.
- The stacking classifier was based on diversity with ML models e.g., RF, XGBoost, SVM thereby reducing the risk of overfitting and improving robustness.



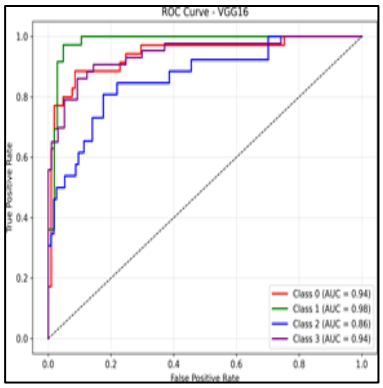
(a) InceptionV3 ROC



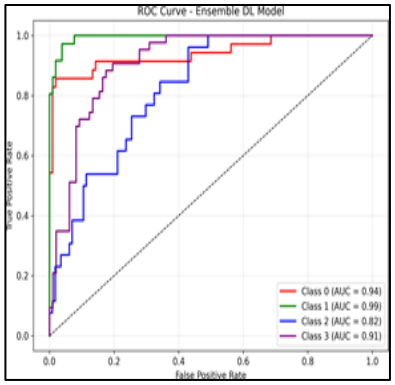
(b) ResNet50 ROC



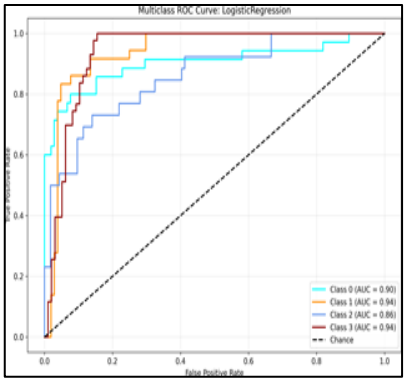
(c) VGG19 ROC



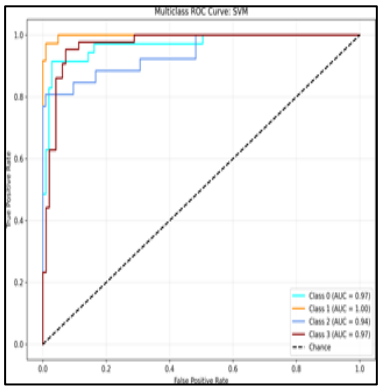
(d) VGG16 ROC



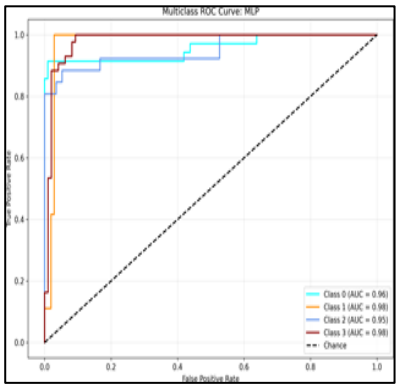
(e) Ensemble DL Model ROC



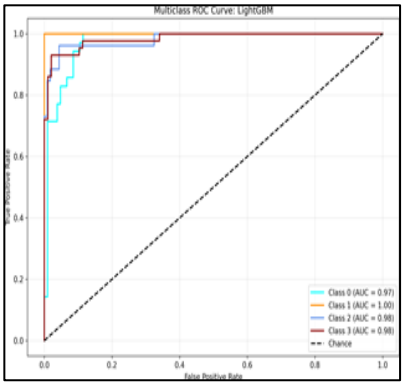
(f) Logistic Regression ROC



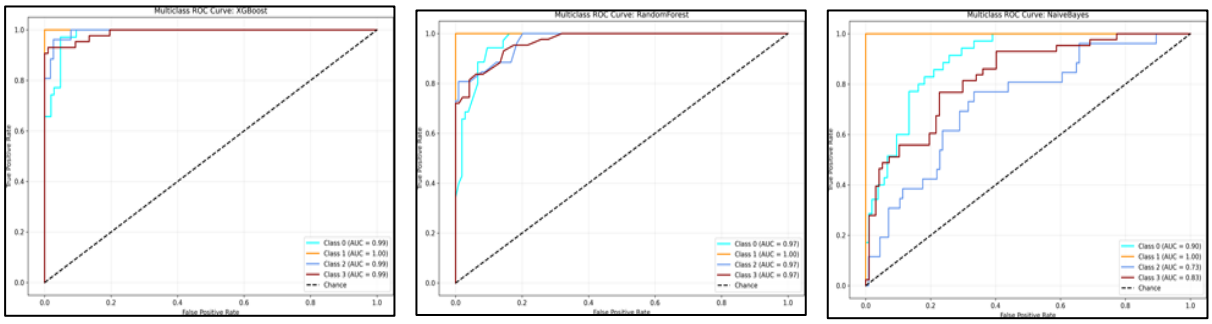
(g) SVM ROC



(h) MLP ROC



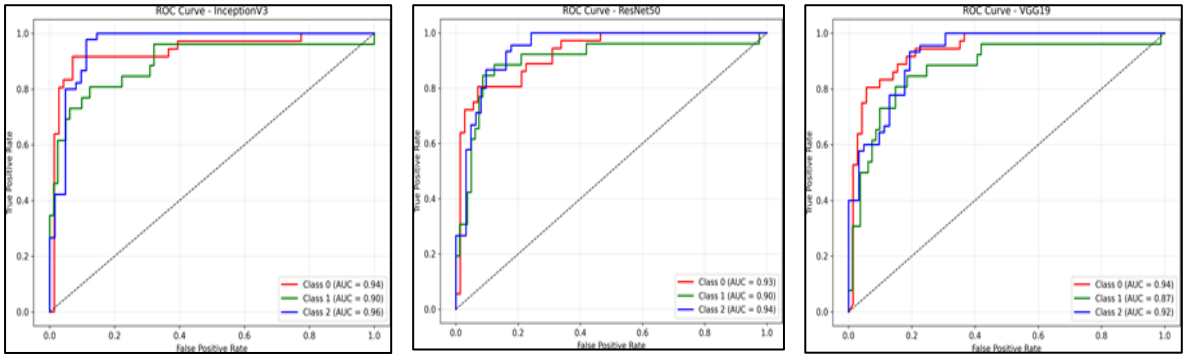
(i) LightGBM ROC



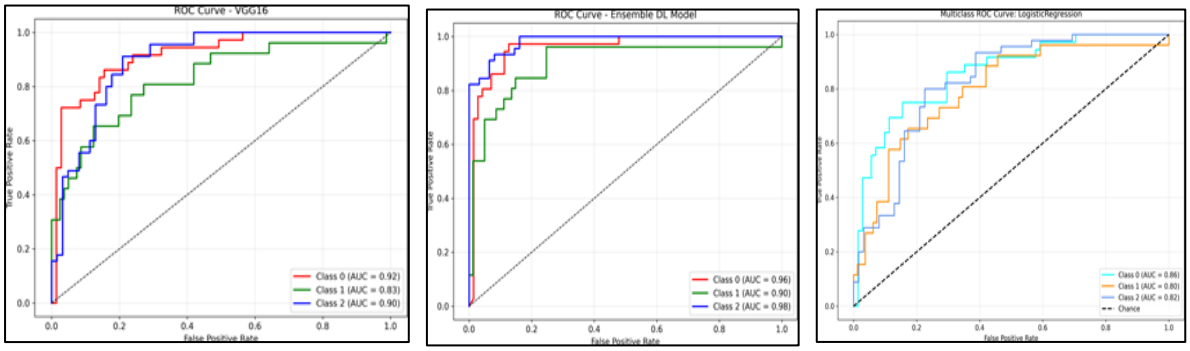
(j) XGBoost ROC (k) Random Forest ROC (l) Naive Bayes ROC

Figure 8. ROC Curves of 12 Models on Dataset 2, Confirming the Ensemble's Superior Classification Performance

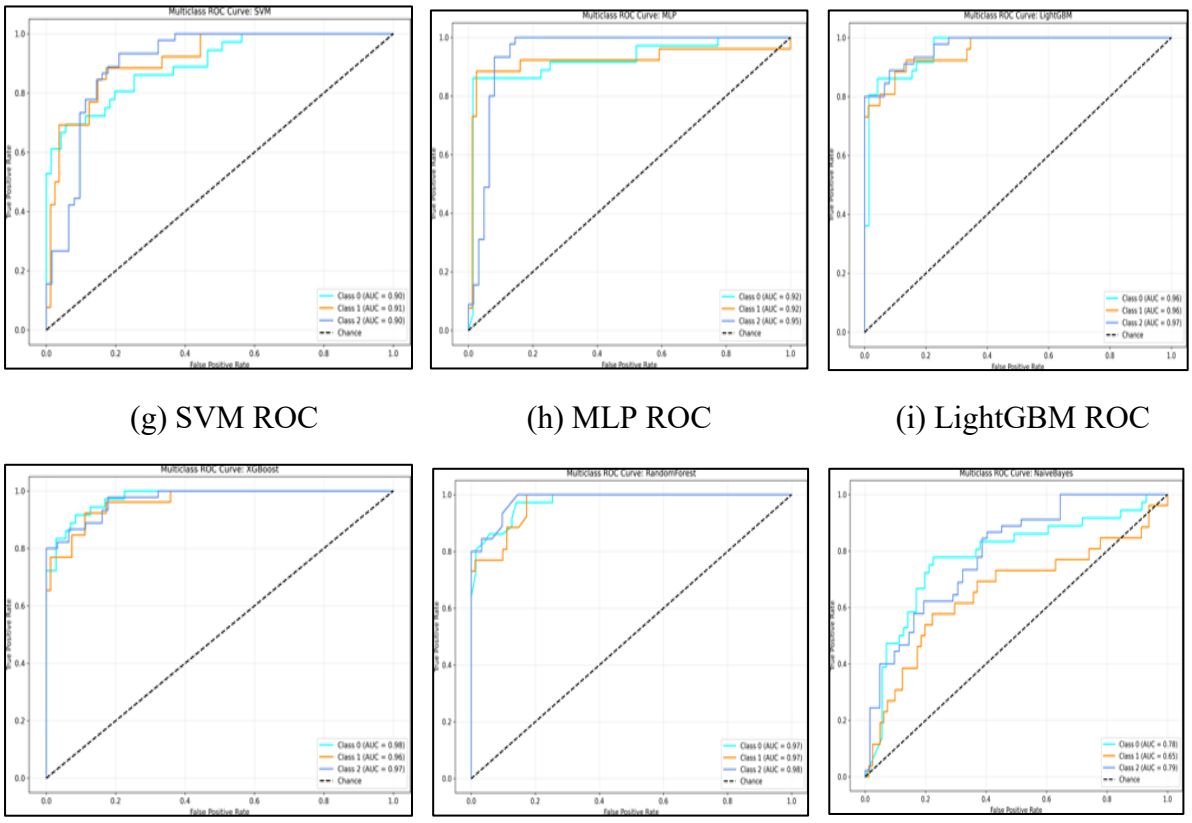
Although you can often achieve slight improvements to a single CNN via deep fine-tuning or hyperparameter tuning, these models are still susceptible to overfitting and biases associated with the particular dataset. Stacking ensembles, on the other hand, are able to use multiple forms of a learner (tree based (RF, XGB, LGBM), kernel based (SVM), neural (MLP)) and can reduce both variance and bias by introducing this heterogeneity. This means stacking ensembles can achieve comparatively, more balanced and generalizable performance regardless of the dataset, a singular optimized CNN lacks the efficiency or protective ability of a stacking ensemble.



(a) InceptionV3 ROC (b) ResNet50 ROC (c) VGG19 ROC



(d) VGG16 ROC (e) Ensemble DL Model ROC (f) Logistic Regression ROC



(g) SVM ROC (h) MLP ROC (i) LightGBM ROC

(j) XGBoost ROC (k) Random Forest ROC (l) Naive Bayes ROC

Figure 9. ROC Curves of 12 Models on Dataset 3, Showing Strong Results Though with Slightly Reduced Separability Between Classes

5.2 Performance Trends Across Datasets

The results obtained from Dataset 1, across all models, were consistently the best, most probably as a result of the size of Dataset 1 and possibly its better labels. This may explain Dataset 2's reasonable performance, while Dataset 3's substantial challenges were likely the result of issues associated with class clarity and the lower representation of classes.

- The ensemble was stable with little deterioration of performance across datasets and showed robustness.
- DL models, on the other hand, lost substantial performance, especially EfficientNet, that consistently under-performed across all datasets.

Overall, the ensemble provided similar predictions across datasets regardless of the size or quality of the labels. It is important to note that Dataset 3 (Mendeley) presented more challenges given the lower number of samples with labeling that was somewhat overlapping with class patterns. However, its performed well in Dataset 3, confirming the ensemble's resilience to dataset scale and labeling differences.

The performance on Dataset 3 was relatively lower likely not only because of its smaller size, but also because of higher class imbalance and less distinct levels of separability of the waveforms between the categories of classes. Notably, the SMOTE augmentation improved recall directly for the minority class. For example, in Dataset 3, the recall for HMI improved by nearly 6% compared to training in the absence of SMOTE. Thus, it would seem that the

SMOTE oversampling strategy was able to mitigate false negatives for clinically important but under-represented classes. Instead of having 4 classes in Datasets 1 and 2, where there was adequate variety of each of the representations, Dataset 3 contained 3 classes and was much smaller. There may have also been inconsistent labeling quality over the recordings, suggesting that future studies with noise-robust training and data augmentation should implement strategies for smaller or under-represented datasets. Upon closer examination of the confusion matrices, we see that the AH and HMI classes show weaker separability, mainly due to overlapping ECG morphologies and slight differences in the waveforms. Therefore, there is a need for richer feature representations or domain-specific augmentation to better separate classes.

5.3 Deep Learning vs Machine Learning Models

CNN-based models like VGG16 and VGG19 produced competitive results with models from Dataset 1, when used individually, they were less effective than traditional ML classifiers. We can state the following observations: DL models are simply better when they have large and diverse datasets for training whereas ML models, specifically tree-based methods and MLP showed adaptability to a total data volume lower than DL, and were substantially benefited from features that were reduced by PCA.

The following points summarize current observations regardless of perceived expectations,

- MLP and XGBoost were consistently the best performers across datasets.
- The DL ensembles yielded limited improvement over the performance of individual DL models, which reinforces the simplicity of MAX based ML decision making frameworks, with a suitable and edges-oriented DL featureing process.

5.4 Confusion Matrix and ROC Observations

The confusion matrices (Figure 6) showed that the proposed ensemble resulted in considerably fewer misclassifications, especially for the more critical classes: Myocardial Infarction (MI), and for the sake of argument, Abnormal Heartbeats (AH). This is particularly important in clinical applications as false negatives could have negative ramifications. The ROC curves again instilled confidence in this approach, with ensemble models exhibiting steeper curve and higher AUC than single models, again providing evidence of better performance as a classifier. That is, we expect to see an improvement in specificity and sensitivity, which is critical to medical diagnostics.

5.5 Practical Implications and Deployment Readiness

The proposed system performed well with data collected from different sources and acquisition settings. The implications of this performance for future, real clinical implementation are strong. For example:

- Noisy ECG data via preprocessing and PCA,
- Class imbalance through SMOTE,

- Multi-class scenarios with stacked classification.

Furthermore, the model we developed has a modular architecture that is well positioned to add new CNN for processes or sites, or to add features for the clinical environment, in order to build a better predictor, faster.

Furthermore, the framework provides lightweight deployability as a consequence of its design: the ECG images are resized to a standard 64×64 greyscale to reduce the input size, PCA reduces the embeddings to 150 dimensions and therefore reduces the computational resource requirements and execution time, and the stacking ensemble only uses inexpensive learners with rapid inference. Such optimizations allow the entire framework to be deployable on mid-range hardware to provide real-time clinical decision support without the need for high computing resources.

Empirical runtime measurements provide further confidence in deployability. Preprocessing was less than 2 ms per image. Feature extraction took 80–100 per sample and the ensemble classifier supplied an additional 5 - 7 ms so rounding, the overall inference time was roughly 100 - 120 ms per ECG image on CPU with memory footprint of < 100 MB with PCA reducing dimensionality by about ∼ 70 % and subsequently memory and compute overhead.

The measured values point to a sufficiently reasonable system for a clinical context to operate in real-time without the need for GPUs on non-GPUs systems. The framework also provides robustness against noise and inter-patient variability through a combination of numerous safeguards: grayscale normalization has been used to minimize illumination-related artifacts, PCA-based dimensionality reduction reduces redundancy that might rely on noise insensitive components, and decision fusion based on ensembles reduces the effect of outlier signals. Together these mechanisms can enforce reliability in real-world ECG data that has common recording discrepancies and physiological variation.

5.6 Limitations

Regardless of the positive findings this study has achieved, there are shortcomings:

- The dataset sizes were relatively small, thus potentially underestimating the variability of ECG patterns within different populations.
- This analysis omitted temporal dynamics, looking at only static snapshots containing ECG images rather than full waveform signals.

5.7 Future Work

Future potential improvements could focus on the following:

- Expanding the framework to allow for larger and multiple lead ECG datasets to improve generalizability and clinical significance.
- Examining time-series models (e.g., LSTM, transformers) for temporal ECG analysis.
- Adding clinical metadata and patient history for context-aware clinical decision-making.

- Investigating lightweight, edge-deployed models for on-device real-time inference on portable devices.

6. Conclusion

The study demonstrates how CNN-based deep feature extraction alongside a stacking ensemble of various ML techniques, provides a more balanced and generalizable solution for ECG image classification than using a single deep learning model. The proposed method reduces overfitting, tackles class imbalance, and maintains performance across heterogeneous datasets. As important as the previous point is, the proposed method makes a leap not just in terms of accuracy, but also in terms of clinical relevance. The method is able to deliver inference in near real-time even on low-end devices; this is critical especially with a view to deploying the method in rural and low-resource healthcare settings. The proposed method further solidifies clinical reliance on AI-enabled cardiovascular diagnosis by improving robustness and scalability. In addition, there seems to be a gap left open for additional research. The proposed method could be enhanced by focusing on multi-lead ECG datasets that are larger in size, applying temporal signal analysis, and implementing edge device or wearable device deployments of lightweight system versions to improve dependability and ease of access. In the context of cardiovascular diseases, the hybrid ML-DL ensemble performs significantly better than the other methods and also emphasizes the importance of intelligent clinical decision support systems for early detection and better management.

References

[1] Naz, Mahwish, Jamal Hussain Shah, Muhammad Attique Khan, Muhammad Sharif, Mudassar Raza, and Robertas Damaševičius. "From ECG signals to images: a transformation based approach for deep learning." PeerJ Computer Science 7 (2021): e386.

[2] Ashtaiwi, AbdulAdhim, Tarek Khalifa, and Omar Alirr. "Enhancing heart disease diagnosis through ECG image vectorization-based classification." Heliyon 10, no. 18 (2024).

[3] Fatema, Kaniz, Sidratul Montaha, Md Awlad Hossen Rony, Sami Azam, Md Zahid Hasan, and Mirjam Jonkman. "A robust framework combining image processing and deep learning hybrid model to classify cardiovascular diseases using a limited number of paper-based complex ECG images." Biomedicines 10, no. 11 (2022): 2835.

[4] Mhamdi, Lotfi, Oussama Dammak, François Cottin, and Imed Ben Dhaou. "Artificial intelligence for cardiac diseases diagnosis and prediction using ECG images on embedded systems." Biomedicines 10, no. 8 (2022): 2013.

[5] Aversano, Lerina, Mario Luca Bernardi, Marta Cimitile, Debora Montano, and Riccardo Pecori. "Early diagnosis of cardiac diseases using ecg images and cnn-2d." Procedia Computer Science 225 (2023): 2866-2875.

[6] Saikumar, Kayam, D. Siva, Dhupati Srivalli, S. Shafiulla Basha, BP Santosh Kumar, and Abolfazl Mehbodniya. "Deep Learning Driven Heart Disease Prediction using ECG Signal Classification." Panamerican Mathematical Journal 35, no. 2s (2025): 42-56.

[7] Sattar, Shoaib, Rafia Mumtaz, Mamoon Qadir, Sadaf Mumtaz, Muhammad Ajmal Khan, Timo De Waele, Eli De Poorter, Ingrid Moerman, and Adnan Shahid. "Cardiac arrhythmia classification using advanced deep learning techniques on digitized ECG

datasets." Sensors 24, no. 8 (2024): 2484.

[8] Narotamo, Hemaxi, Mariana Dias, Ricardo Santos, André V. Carreiro, Hugo Gamboa, and Margarida Silveira. "Deep learning for ECG classification: A comparative study of 1D and 2D representations and multimodal fusion approaches." Biomedical Signal Processing and Control 93 (2024): 106141.

[9] Ayano, Yehualashet Megersa, Friedhelm Schwenker, Bisrat Derebssa Dufera, Taye Girma Debelee, and Yitagesu Getachew Ejegu. "Interpretable hybrid multichannel deep learning model for heart disease classification using 12-lead ECG signal." IEEE Access 12 (2024): 94055-94080.

[10] Khan, Hira, Nadeem Javaid, Tariq Bashir, Mariam Akbar, Nabil Alrajeh, and Sheraz Aslam. "Heart disease prediction using novel ensemble and blending based cardiovascular disease detection networks: EnsCVDD-Net and BlCVDD-Net." IEEE Access 12 (2024): 109230-109254.

[11] Nawaz, Syed Ali, Muhammad Arshad, Tanveer Aslam, Abdul Haseeb Wajid, Rizwan Ali Shah, and Mubashir H. Malik. "ECG based heart disease diagnosis using machine learning approaches." Journal of Computing & Biomedical Informatics (2024).

[12] Gajendran, Mohan Kumar, Muhammad Zubair Khan, and Muazzam A. Khan Khattak. "Ecg classification using deep transfer learning." In 2021 4th international conference on information and computer technologies (ICICT), IEEE, (2021): 1-5.

[13] Sinha, Nidhi, MA Ganesh Kumar, Amit M. Joshi, and Linga Reddy Cenkeramaddi. "DASMcC: Data Augmented SMOTE Multi-class Classifier for prediction of Cardiovascular Diseases using time series features." IEEE Access 11 (2023): 117643-117655.

[14] Huang, Pang-Shuo, Yu-Heng Tseng, Chin-Feng Tsai, Jien-Jiun Chen, Shao-Chi Yang, Fu-Chun Chiu, Zheng-Wei Chen et al. "An artificial intelligence-enabled ECG algorithm for the prediction and localization of angiography-proven coronary artery disease." Biomedicines 10, no. 2 (2022): 394.

[15] Karthik, S., M. Santhosh, Muthu Subash Kavitha, and A. Christopher Paul. "Automated Deep Learning Based Cardiovascular Disease Diagnosis Using ECG Signals." Computer Systems Science & Engineering 42, no. 1 (2022).

[16] Mishra, Jyoti, and Mahendra Tiwari. "IoT-enabled ECG-based heart disease prediction using three-layer deep learning and meta-heuristic approach." Signal, Image and Video Processing 18, no. 1 (2024): 361-367.

[17] Yoon, Taeyoung, and Daesung Kang. "Multi-modal stacking ensemble for the diagnosis of cardiovascular diseases." Journal of Personalized Medicine 13, no. 2 (2023): 373.

[18] Mahmud, Tanjim, Anik Barua, Dilshad Islam, Mohammad Shahadat Hossain, Rishita Chakma, Koushick Barua, Mahabuba Monju, and Karl Andersson. "Ensemble deep learning approach for ecg-based cardiac disease detection: Signal and image analysis." In 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), IEEE, (2023): 70-74.

[19] Alsekait, Deema Mohammed, Ahmed Younes Shdefat, Ayman Nabil, Asif Nawaz, Muhammad Rizwan Rashid Rana, Zohair Ahmed, Hanaa Fathi, and Diaa Salama AbdElminaam. "Heart-Net: A Multi-Modal Deep Learning Approach for Diagnosing Cardiovascular Diseases." Computers, Materials & Continua 80, no. 3 (2024).

[20] Dhara, Sanjib Kumar, Nilankar Bhanja, and Prabodh Khampariya. "An adaptive heart disease diagnosis via ECG signal analysis with deep feature extraction and enhanced

radial basis function." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 11, no. 7 (2024): 2245927.

[21] Ma, Linjuan, and Fuquan Zhang. "A novel real-time detection and classification method for ECG signal images based on deep learning." Sensors 24, no. 16 (2024): 5087.

[22] Akter, M., Islam, N., Ahad, A., Chowdhury, M. A., Apurba, F. F., & Khan, R. (2024). An embedded system for real-time atrial fibrillation diagnosis using a multimodal approach to ECG data. Eng, 5(4), 2728–2751. https://doi.org/10.3390/eng5040143

[23] Alsayat, A., Mahmoud, A. A., Alanazi, S., Mostafa, A. M., Alshammari, N., Alrowaily, M. A., Shabana, H., & Ezz, M. (2025). Enhancing cardiac diagnostics: A deep learning ensemble approach for precise ECG image classification. Journal of Big Data, 12(1), 7. https://doi.org/10.1186/s40537-025-01070-4

[24] Mohsin, K. (2022). National Heart Foundation 2023 ECG dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/drkhaledmohsin/national-heart-foundation-2023-ecg-dataset

[25] Khan, A. H., & Hussain, M. (2021). ECG images dataset of cardiac patients (Version 2) [Data set]. Mendeley Data. https://doi.org/10.17632/gwbz3fsgp8.2

[26] Ray, A. (2024). ECG dataset for heart condition classification (Version 2) [Data set]. Mendeley Data. https://doi.org/10.17632/xw9sd3btcs.z

[27] Mahajan, Palak, Shahadat Uddin, Farshid Hajati, and Mohammad Ali Moni. "Ensemble learning for disease prediction: A review." In Healthcare, vol. 11, no. 12, MDPI, (2023): 1808.

[28] Salman, Hasan Ahmed, Ali Kalakech, and Amani Steiti. "Random forest algorithm overview." Babylonian Journal of Machine Learning 2024 (2024): 69-79.

[29] Zhang, Ping, Yiqiao Jia, and Youlin Shang. "Research and application of XGBoost in imbalanced data." International Journal of Distributed Sensor Networks 18, no. 6 (2022): 15501329221106935.

[30] Al Bataineh, Ali, Devinder Kaur, and Seyed Mohammad J. Jalali. "Multi-layer perceptron training optimization using nature inspired computing." IEEE Access 10 (2022): 36963-36977.

[31] Roy, Atin, and Subrata Chakraborty. "Support vector machine in structural reliability analysis: A review." Reliability Engineering & System Safety 233 (2023): 109126.

[32] Schober, Patrick, and Thomas R. Vetter. "Logistic regression in medical research." Anesthesia & Analgesia 132, no. 2 (2021): 365-366.

[33] Zabor, Emily C., Chandana A. Reddy, Rahul D. Tendulkar, and Sujata Patil. "Logistic regression in clinical studies." International Journal of Radiation Oncology* Biology* Physics 112, no. 2 (2022): 271-277.