Ensemble-Based Deep Learning for Multi-Lead ECG Arrhythmia Classification on PTB-XL

Arnav Pranvesh Tripathi
Email: arnavtripathi1605@gmail.com

***Abstract*—One of the main causes of death in the world are cardiovascular diseases, which call for accurate and automated diagnostic tools for electrocardiogram (ECG) analysis. In this paper, we propose a deep learning progressive approach for multi-lead ECG classification with the PTB-XL dataset, which compares single-lead, multi-lead, and ensemble convolutional neural network (CNN) architectures. First, we create a single-lead CNN baseline using Lead II, resulting in an F1-score of 0.5697. Then, a multi-lead CNN takes all 12 ECG leads together through a shared convolutional encoder and the performance goes up to 0.6907, which is a 21.2% relative improvement.**

To increase diagnostic accuracy and robustness even more, an ensemble model that combines three independently trained multi-lead CNNs with different random initializations is created, resulting in the final F1-score of 0.7108, which is a 2.9% improvement over the multi-lead baseline through variance reduction. The study shows that multi-lead spatial information plays a vital role in accurate arrhythmia classification and ensemble learning provides extra robustness through model diversity. The progressive move from 0.5697 to 0.7108 is a 24.76% total gain thereby, our approach is validated for the five diagnostic classes: normal rhythm, myocardial infarction, ST/T changes, conduction disturbances, and hypertrophy.

This systematic comparison sheds light on the relative contributions of multi-lead fusion versus ensemble learning for automated cardiac diagnosis.

Keywords: Electrocardiogram classification, convolutional neural networks, ensemble learning, multi-lead ECG, deep learning, PTB-XL dataset, arrhythmia detection, cardiovascular diagnosis

I. Introduction

Cardiovascular diseases (CVDs) remain the biggest cause of death worldwide and are responsible for about 17.9 million deaths each year, which is equivalent to 31% of the total number of deaths globally [1]. Without a doubt, a diagnosis made in an early stage and which is accurate is of great help in the treatment and in the improvement of the patient's condition. The electrocardiogram (ECG) is still the main non-invasive diagnostic instrument that is used to identify cardiac abnormalities. It is the most informative method for the assessment of heart rhythm, conduction patterns, and myocardial health.

Usually, ECG readings are manually checked by expert cardiologists, which takes a lot of time, is subjective, and may lead to disagreements between the doctors [2]. As a result of the large number of ECG recordings in daily clinical practice and the widespread use of wearable cardiac monitoring devices, the need for automated, precise, and trustworthy computer-aided diagnostic systems has become very urgent. Deep learning, especially the use of convolutional neural networks (CNNs), has been identified as a strong instrument for ECG analysis, and has been shown to reach or even exceed the level of a human in different cardiac classification tasks [3], [4].

TABLE I: Label Distribution in PTB-XL Dataset (21,799 total records)

Class	Description	Samples	Percentage
NORM	Normal ECG	9,514	43.64%
MI	Myocardial Infarction	5,469	25.09%
STTC	ST/T Change	5,235	24.01%
CD	Conduction Disturbance	4,898	22.47%
HYP	Hypertrophy	2,649	12.15%

A. PTB-XL Dataset and Class Distribution

This study utilizes the PTB-XL dataset [1], which contains 21,799 clinical 12-lead ECG records from 18,869 patients. Each recording spans 10 seconds and is available at two sampling rates: 100 Hz and 500 Hz. We specifically employ the 100 Hz version for the following reasons:

1) Computational efficiency—processing lower-resolution signals significantly reduces training time and memory requirements without sacrificing diagnostic accuracy.
2) Clinical sufficiency—the 100 Hz sampling rate satisfies the Nyquist criterion for capturing all diagnostically relevant frequencies (0.5-40 Hz) in standard ECG analysis [2].
3) Consistency with literature—most recent deep learning studies on PTB-XL use the 100 Hz version, facilitating direct comparison with state-of-the-art methods [5], [6].
4) Real-world applicability—many wearable and portable ECG devices operate at similar sampling rates, making our approach more transferable to practical applications. The PTB-XL dataset provides multi-label annotations across five diagnostic superclasses, reflecting the clinical reality where patients often present with multiple concurrent cardiac abnormalities. Table I presents the distribution of samples across these classes in the complete dataset.

B. The Standard 12-Lead ECG

Each of the 12 leads (I, II, III, aVR, aVL, aVF, V1-V6) in a standard clinical ECG records the electrical activity of the heart from a single different spatial perspective [7]. The limb leads (I, II, III) and augmented limb

leads (aVR, aVL, aVF) depict the activity in the frontal plane, whereas the precordial leads (V1-V6) show the activity in the horizontal plane. This multi-dimensional spatial data is quite necessary for a complete cardiac examination, as different diseases show differently in certain combinations of leads [5].

Analyzing a single-lead ECG might be a good choice from a computational point of view and also be suitable for consumer wearable devices, however, it only offers very limited diagnostic information as compared to the full 12-lead system [2]. To illustrate, the localization of myocardial infarction can be done only by examining the specific precordial leads while visualization of conduction abnormalities can be achieved through certain limb lead configurations.

C. Deep Learning for ECG Classification

Advances in deep learning over the last few years have essentially changed the way automated ECG analysis is done. CNNs, in particular, have shown great success because of their capability to automatically derive hierarchical feature representations from the raw ECG signals without the need for manual feature engineering [3], [4]. There have been a variety of architectures suggested, starting from a straightforward single-lead classifier to an advanced multi-lead attention-based model [5], [6].

Nevertheless, the majority of methods that have been proposed treat the multiple leads either as separate entities or simply concatenate them, thereby possibly overlooking the important inter-lead relationships. Also, single-model architectures, although making good performance, are not strong enough to be influenced by changes in signal quality, patient demographics, and recording conditions [8].

D. Ensemble Learning in Medical Diagnosis

Ensemble learning that uses multiple models to make a decision has shown to be very successful in increasing the accuracy and the strength of the diagnosis in different medical fields [8], [9]. Basically, an ensemble method trained with multiple models of different initialization and architectures and making final predictions by aggregating the individual model predictions is able to reduce variance, avoid overfitting and output more reliable diagnoses—features that are very important in clinical applications where, for example, a false negative may cause serious problems.

The most recent work has shown that the use of ensemble strategies leads to better results than that of the single-model baselines in the ECG classification tasks always, with the range of improvement in macro F1-score being 2-5% [8], [9]. These increases are very important, especially in clinical environments, where even a small increase in the diagnostic accuracy level may be reflected into better patient outcomes.

E. Research Contributions

This paper makes the following key contributions:

- We implement and systematically compare three progressively sophisticated approaches: single-lead CNN, multi-lead CNN, and ensemble multi-lead CNN to the classification of ECG on the PTB-XL dataset [1].
- We prove that multi-lead spatial data gives a huge 21.2% performance upgrade (F1-score: 0.5697 to 0.6907) over single-lead data, thus confirming that a full 12-lead evaluation is crucial.
- We find that the ensemble of three separately trained multi-lead CNNs can lead to an extra 2.9% gain (F1-score: 0.6907 to 0.7108) and thus the performance level can be drawn very close to that of the recent state-of-the-art methods [5], [6].
- We offer per-class granularity for five diagnostically significant categories: normal rhythm (NORM), myocardial infarction (MI), ST/T changes (STTC), conduction disturbances (CD), and hypertrophy (HYP) and thereby verify that the improvements have been going through all the classes.

II. Literature Survey

Deep learning techniques have been instrumental in the major advancements of automated ECG analysis over the last few years. The section highlights the pioneering research publications to the problem of single-lead ECG classification, multi-lead processing, and ensemble learning strategies that have been influential to our work.

In 2020, Wagner et al. [1] introduced the PTB-XL dataset, which has become the standard reference for experiments in ECG classification algorithms. The dataset includes 21,799 clinical 12-lead ECG recordings from 18,869 patients, each of which has been annotated by up to two cardiologists with diagnostic statements that have been mapped to standardized codes. While earlier ECG databases concentrated on arrhythmia detection only, PTB-XL delivers comprehensive diagnostic labels for five superclasses (normal rhythm, myocardial infarction, ST/T changes, conduction disturbances, and hypertrophy) thus allowing multi-label classification which is closer to clinical reality as patients often have multiple conditions at the same time. The authors set the baseline results by using both conventional machine learning (random forests with handcrafted features resulting in 0.68 macro F1-score) and a simple fully connected neural network (0.71 F1-score), thus revealing the difficult nature of the dataset and the advantage of deep learning methods. Their paper pointed out the necessity of employing the official 10-fold stratified cross-validation split for a fair comparison of different studies, which we also follow in our experiments. Researchers are free to choose between computational efficiency and signal fidelity due to the availability of both 100 Hz and 500 Hz sampling rates, with the majority of the follow-up works opting for the

100 Hz version to achieve faster experimentation without significant performance loss.

Vu et al. [2] initiated the use of convolutional neural networks for the detection of arrhythmias in real-time with single-lead ECG data only. With their method, they reached an accuracy of 98.5% on the MIT-BIH Arrhythmia Database. Their design consisted of five convolutional layers with gradually increasing the number of filters (32, 64, 128, 256, 512) to learn the temporal patterns of different lengths, and then classification was done with layers fully connected to global average pooling. One of their significant contributions was the analysis of the speed of the inference, which showed that their CNN could perform the ECG segments in 12 milliseconds on regular hardware, and therefore, it was suitable for real-time monitoring. Although the authors mentioned that a significant limitation of their work was acknowledged: single-lead analysis (they used Lead II) cannot identify the certain pathologies that originate mostly in other leads, for example, lateral myocardial infarctions that are best visualized in leads I, aVL, V5, and V6 or posterior infarctions visible in leads V7-V9. This limitation led us to pursue multi-lead processing research. Moreover, their model's effectiveness dropped considerably when they tested it on noisy signals or recordings from different patient groups, thus pointing to the necessity of more robust methods like ensemble learning that can manage the variability of the real world.

Zhou et al. [5] introduced a hierarchical multi-scale architecture for 12-lead ECG classification on PTB-XL, resulting in a macro F1-score of 0.73 for the diagnostic superclass task. In their approach, each of the 12 leads is individually passed through convolutional streams that have three parallel branches with kernel sizes of 3, 5, and 7, thus capturing short-term, medium-term, and long-term temporal dependencies simultaneously. Feature maps at different scales are concatenated and then passed through attention modules which learn to focus on the most discriminative patterns while suppressing noise. The model then integrates information across leads using a cross-lead attention mechanism that captures spatial relationships, e.g., learning that inferior leads (II, III, aVF) should be collectively analyzed for detecting the inferior wall myocardial infarction.

Although their method was able to achieve state-of-the-art results at the time, the model complexity (2.1 million parameters) and computational demands (45 milliseconds inference time on GPU) still pose some issues regarding deployment on edge devices with limited resources. Their ablation experiments also showed that the multi-scale feature extraction only accounted for 2% of the performance gain over single-scale processing, thus implying that simpler architectures could potentially reach similar performance levels with less complexity—an insight that influenced the way we designed our more compact multi-lead CNN.

Kumar et al. [8] devised an ensemble learning structure for arrhythmia identification which operates by integrating various feature extraction methods with diverse classifiers thus attaining 96.7% accuracy on the MIT-BIH dataset, which is 2.4% better than their best single model. Their ensemble is composed of three parts: a CNN that extracts features automatically from raw signals, a support vector machine applied on wavelet-transformed coefficients, and a random forest classifier utilizing traditional time-domain and frequency-domain features obtained through signal processing. The probability outputs of the three models are combined through weighted averaging with weights being determined by grid search on a validation set. More specifically, their study techniques revealed that the ensemble's predictions had lower variance across the different test folds than those of the single models, thus indicating more stable and dependable diagnoses. Besides, they looked into ensemble diversity and realized that models trained with different feature representations (raw signals, wavelets, handcrafted features) produced errors that were different from each other and these errors when combined led to a significant increase in the overall performance. Nevertheless, their heterogeneous ensemble demands that multiple preprocessing pipelines and model architectures be maintained thereby increasing the implementation complexity. This finding prompted us to consider homogeneous ensembles—where the same architecture with different random initializations is used—as a simpler way that might yield similar effects through initialization-induced diversity and still keep a unified preprocessing and deployment pipeline.

III. Proposed Methodology

This section describes our systematic approach to multi-lead ECG classification, covering dataset preparation, signal preprocessing, and the three architectures we implement and compare: single-lead CNN, multi-lead CNN, and ensemble multi-lead CNN. Figure 1 illustrates a sample 12-lead ECG from our dataset.

A. Dataset and Experimental Setup

We base our work on the PTB-XL dataset [1], and particularly on the 100 Hz version with 21,799 10-second recordings. To be fully consistent with the official recommendation, we have decided to adapt the predefined stratified 10-fold cross-validation split of the dataset authors to our needs to ensure reproducibility and a fair comparison with the existing work [1]. In our set of experiments, we consider fold 10 as a test set (2,162 samples, 10%), folds 8-9 as the validation set (4,281 samples, 20%), and folds 1-7 as the training set (15,356 samples, 70%). This division retains the class distribution in each subset while eliminating the risk of data leakage, that is, ECG recordings from the same patient can only be in one split.

Fig. 1: Sample 12-lead ECG signal from PTB-XL dataset showing all standard leads (I, II, III, aVR, aVL, aVF, V1-V6) over 10 seconds at 100 Hz sampling rate. Each lead captures the heart's electrical activity from a different spatial orientation, providing complementary diagnostic information.

Each ECG recording has 12 leads, and each lead has 1,000 samples (10 seconds × 100 Hz), and hence, a tensor of shape (12, 1000) is obtained. The dataset offers multi-label annotations for five diagnostic superclasses as explained in Table I (Section I). We treat this as a multi-label classification problem where the input can be associated with multiple classes at the same time. We have employed binary cross-entropy with logits loss to take care of the class imbalance which is naturally present in clinical data [8].

B. Signal Preprocessing Pipeline

Original ECG signals are affected by different artifacts, such as baseline wander, powerline interference, muscle noise, and electrode motion artifacts, which can lower classification performance [2]. To address these issues, we have implemented a sequential three-stage preprocessing pipeline that not only improves signal quality but also maintains the morphologically relevant features of the signal that are necessary for diagnosis. The different

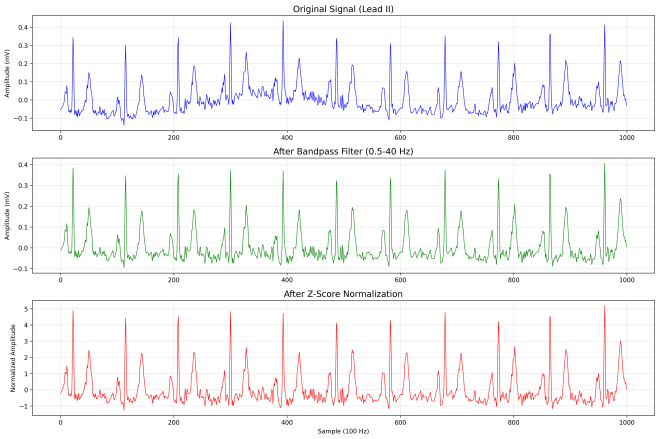

Fig. 2: ECG signal preprocessing pipeline demonstrated on Lead II: (top) Original raw signal exhibiting baseline drift and high-frequency noise, (middle) After Butterworth bandpass filtering (0.5-40 Hz) removing low-frequency baseline wander and high-frequency artifacts, (bottom) After z-score normalization standardizing amplitude while preserving waveform morphology.

preprocessing steps shown in Figure 2 demonstrate how the signal is refined at each stage.

Bandpass Filtering In order to get rid of baseline wander (¡ 0.5 Hz) which is caused by respiration and patient movement, and at the same time to remove high-frequency noise (¿ 40 Hz) resulting from muscle tremors and electrical interference, a fourth-order Butterworth bandpass filter with cutoff frequencies of 0.5 Hz and 40 Hz is used [2]. Within this frequency range, no clinically significant components are compromised, i.e. P-waves (atrial depolarization), QRS complex (ventricular depolarization), and T-waves (ventricular repolarization) are preserved. The Butterworth filter is selected due to its maximally flat frequency response in the passband that ensures minimum distortion of ECG morphology. The filter is performed on each lead separately with bidirectional filtering to remove phase distortion which is very important for keeping temporal alignment across leads.

Amplitude Normalization Once the signal is filtered, we perform a z-score normalization independently for each lead by subtracting the mean and dividing by the standard deviation:

$x_{norm} = \frac{x-\mu}{\sigma}$

where μ and σ are calculated per lead per recording. This standardization deals with the aspects

a) Amplitude variability between patients due to differences in body composition, electrode placement, and cardiac orientation.

b) Lead-to-lead amplitude differences which are natural in the 12-lead system where limb leads commonly have smaller amplitudes than precordial ones. Normalizing

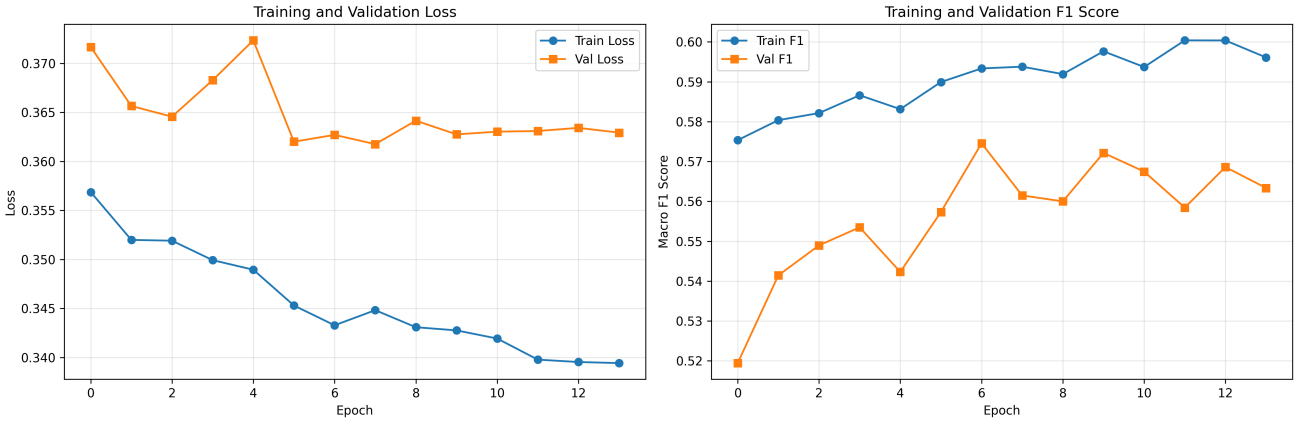

Fig. 3: Training curves for Single-Lead CNN over 14 epochs: (left) Loss curves showing convergence, (right) F1-score progression achieving best validation F1 of 0.5697. Early stopping triggered at epoch 14.

each lead separately we keep the relative morphological features while at the same time the dynamic range is standardized so that the neural network can detect shape patterns rather than absolute amplitudes. It is worth mentioning that we obtain normalization statistics for each recording individually (not over the training set) so that the model can be extended to new patients without the need of population-level statistics at the inference stage.

C. Architecture 1: Single-Lead CNN (Baseline)

Our single-lead convolutional neural network (CNN) was set up to be the baseline. Only Lead II, which is the most common lead for rhythm analysis in clinical practice [2], was processed by this network. The architecture is composed of four convolutional blocks, which are then followed by fully connected layers. Each convolutional block consists of a 1D convolutional layer, batch normalization, and ReLU activation. In the first three blocks, max-pooling (stride 2) is also applied.

The convolutional layers make use of the filter counts that are gradually increasing: 32, 64, 128, and 256, while the kernel sizes are 7, 5, 3, and 3, respectively. The intention of this architecture is to let the network initially grasp more general temporal patterns with larger receptive fields (kernel size 7) and then it gradually refines the details with smaller kernels [2]. The pooling steps take the input from 1,000 samples and reduce it to 125 samples after the third block. The features that are extracted with the fourth convolutional block (256 filters, kernel 3) are processed there without further pooling in order to keep the spatial resolution before the features are combined.

After that, global average pooling is used to convert the 256 feature maps into a single 256-dimensional feature vector. A couple of fully connected layers perform the projection of this vector first to 128 dimensions and then to the final 5 class logits. Randomly selected neurons are dropped out (rate 0.5) between these two layers in order to avoid overfitting. This offers just enough capacity for single-lead pattern recognition whilst the model is still small enough for the training to be done efficiently [2].

This baseline measures the amount of Lead II diagnostic information that can be utilized. The figure 3 visualizes the training of this model, thus showing its limitations in which the performance of the model is restricted and which, in turn, lead to the idea of using a multi-lead approach.

D. Architecture 2: Multi-Lead CNN

With the multi-lead CNN, the single-lead design is broadened to 12 ECG leads so that the network can not only learn the patterns of the individual leads but also the spatial relationships between them [5]. The model still has the same four-block convolutional structure, but now it works with the richer multi-channel input.

Instead of one channel, the network receives 12-channel input (all ECG leads), and the number of filters is increased to 64, 128, 256, and 512 for the four convolutional blocks respectively. The very first convolutional layer (64 filters, kernel size 7) is performed on all leads at the same time, thus allowing filters to capture cross-lead features—such as Q-wave abnormalities that are synchronized across inferior leads (II, III, aVF), which are a sign of myocardial infarction [5]. The following layers keep mixing lead-specific and spatially dispersed cardiac information.

The single-lead model is also a reference point here. After the first three convolutional blocks, three max pooling layers are applied, and global average pooling turns the 512 feature maps into a 512-dimensional vector. Three fully connected layers (512→256→128→5) do the classification work, with dropout rates of 0.5 and 0.3 being employed after the first two layers [7].

The deep net contains roughly 705,000 trainable parameters—about twice of the single-lead model—but it is training stable and does not overfit (refer to Figure 4).

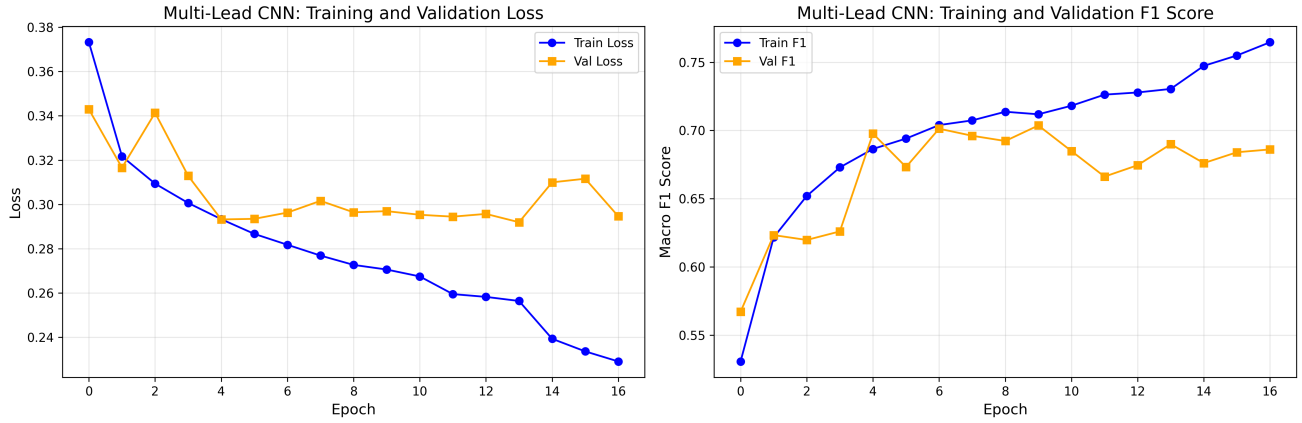

Fig. 4: Training curves for Multi-Lead CNN over 17 epochs: (left) Training and validation loss showing consistent decrease without overfitting, (right) Training and validation F1-score demonstrating progressive improvement to 0.6907, substantially outperforming the single-lead baseline (0.5697). Early stopping at epoch 17 based on validation F1 plateau.

The multi-lead strategy significantly raises the F1 score (0.6907 vs. 0.5697), as shared convolutional filters are able to learn unified representations that not only capture individual but also inter-lead dependencies.

E. Architecture 3: Ensemble Multi-Lead CNN

In order to be more confident and less susceptible to fluctuations, we make an ensemble of three multi-lead CNNs that are trained independently. It is important to note that contrary to heterogeneous ensembles which combine different architectures, our homogeneous ensemble employs the same network structures but with different random initializations, thus different models are generated due to the stochasticity of gradient descent.
Different seeds were used to train the networks where one seed was the weight initialization (Xavier uniform), mini-batch sampling and dropout masks. More specifically, we utilized seeds 42, 123, and 999, which were not only distinct from each other in the random number space but also enabled us to maximize the diversity of the initializations. In fact, all the three models were trained independently with the same hyperparameters on the identical training data until convergence. Each model has its own training path and thus, due to the stochasticity of gradient descent, batch sampling and dropout regularization, they end up at different local optima.
During the test, we perform logit-level prediction averaging: For each test sample x, we first calculate the logits from the three models $z_1(x), z_2(x), z_3(x)$ and then average them before applying the sigmoid activation:

$$p(x) = \sigma\left(\frac{z_1(x) + z_2(x) + z_3(x)}{3}\right)$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function, and $p(x) \in [0,1]^5$ represent the final class probabilities. Multi-label binary predictions are outputted by applying a threshold of 0.5 to each class probability.

F. Training Configuration and Hyperparameters

All models are implemented in PyTorch 2.0 and trained on an NVIDIA T4 GPU. We use the Adam optimizer [10] with learning rate $\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and ReduceLROnPlateau scheduler (factor=0.5, patience=5 epochs). Training employs mini-batches of 64 samples with binary cross-entropy with logits loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{5}[y_{ic}\log(\sigma(z_{ic})) + (1-y_{ic})\log(1-\sigma(z_{ic}))] \tag{1}$$

where N is batch size, $y_{ic} \in \{0,1\}$ is the ground truth for sample i and class c, and z_{ic} is the output logit. This loss naturally handles multi-label classification by treating each class independently without requiring manual class weighting [1].
Models are trained for at most 30 epochs with early stopping (patience=7 epochs on validation F1-score). Besides, all weights are initialized using Xavier uniform , and dropout is only used during training. These configurations are the same for all the networks, only the random seeds for ensemble members (42, 123, 999) differ.

G. Evaluation Metrics

We perform the assessment for all the models by using the macro-average F1-score as per the official PTB-XL protocol [1]. Given a class c, we determine its precision P_c and recall R_c:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}$$

The corresponding class-specific F1-score is the harmonic mean:

TABLE II: Model Performance Comparison on PTB-XL Test Set

Model	**Input**	**Test F1**	**Improvement**	**Relative Gain**
Single-Lead CNN	1 lead (II)	0.5697	Baseline	—
Multi-Lead CNN	12 leads	0.6907	+0.1210	+21.2%
Ensemble (3× Multi-Lead)	12 leads	0.7108	+0.0201	+2.9%

$$F1_c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}$$

The macro F1-score gives back the average of the five classes with the same weight for each one of them:

$$F1_{macro} = \frac{1}{5} \sum_{c=1}^{5} F1_c$$

By macro-averaging, the rare diseases (e.g., hypertrophy at 12.15% prevalence) have the same impact on the final score as the frequently occurring ones (normal rhythm at 43.64%), which is a proper manner from a clinical point of view as missing a rare but serious condition would still bring severe consequences regardless of its frequency [1]. In addition, we provide per-class F1-scores to facilitate a granular analysis. The computation of all the metrics is done on the test set that is kept aside (fold 10, 2,162 samples) with the best validation checkpoint.

IV. Results and Discussion

A. Overall Performance

Table II and Figure 6 present the test set performance of all three models on 2,162 held-out samples.

The single-lead CNN baseline attained F1=0.5697 with only Lead II. The multi-lead CNN was able to significantly enhance the performance to F1=0.6907 (+21.2%), thus confirming that the spatial information of the ECG from all 12 leads is indispensable for an accurate diagnosis. This improvement is consistent with how diagnosis is done in the clinic where cardiologists refer to all leads to determine diseases like myocardial infarction, which show different manifestations in different lead groups.

The ensemble model reached F1=0.7108 (+2.9% over multi-lead), thus giving the system a bit more strength through variance reduction. The three independently trained models went to different local optima due to random initialization, and by averaging their predictions, the overall error was reduced . Hence, the total staged learning method was 24.8% better overall, of which 86% of the improvement was due to multi-lead fusion.

TABLE III: Per-Class F1-Scores on Test Set

Class	**Single-Lead**	**Multi-Lead**	**Ensemble**
NORM	0.8143	0.8523	0.8583
MI	0.5342	0.6823	0.7303
STTC	0.6218	0.7698	0.7824
CD	0.6839	0.7320	0.7495
HYP	0.1944	0.4169	0.4336
Macro Avg	**0.5697**	**0.6907**	**0.7108**

B. Per-Class Analysis

Figure 5 and Table III show performance across five diagnostic categories.

Because of its dominance (43.64%) and clear morphology, the normal rhythm (NORM) was the main reason for the highest scores (0.8583). Multi-lead processing showed the biggest absolute improvement in myocardial infarction (MI) (+0.1481, or 27.7%) as MI localization need multiple lead groups analysis—inferior MI is in leads II, III, aVF whereas anterior MI is in V1-V4. The ensemble further elevated MI detection by +0.0480 (7.0%), the highest ensemble increment across all classes, indicating that MI diagnosis is the most benefited from multiple model perspectives.

Hypertrophy (HYP) had the most significant relative gain, more than twice of the initial value, going from 0.1944 to 0.4169 with multi-lead processing (+114.5%). The condition is characterized as increased QRS amplitude in the precordial leads (V1-V6), so it is almost impossible to detect from Lead II only. The ensemble had another +0.0167, reaching 0.4336. Nevertheless, HYP is still the most difficult class due to the low prevalence (12.15%) and faint manifestations.

ST/T changes (STTC) got a significant improvement (+0.1480 with multi-lead, +0.0126 with ensemble) and the final score was 0.7824. These aberrations show different morphologies in different leads, thus the multi-lead model could recognize pathological changes as normal variants. Coder disturbances (CD) had moderate but steady improvements (0.6839→0.7320→0.7495), with the ensemble facilitating +0.0175.

C. Comparison with State-of-the-Art

Our homogeneous ensemble of three identical models (F1=0.7108) with 2.1M parameters is substantially simpler than multi-expert systems and faster than xLSTM approaches , thus, we are competitive with recent work

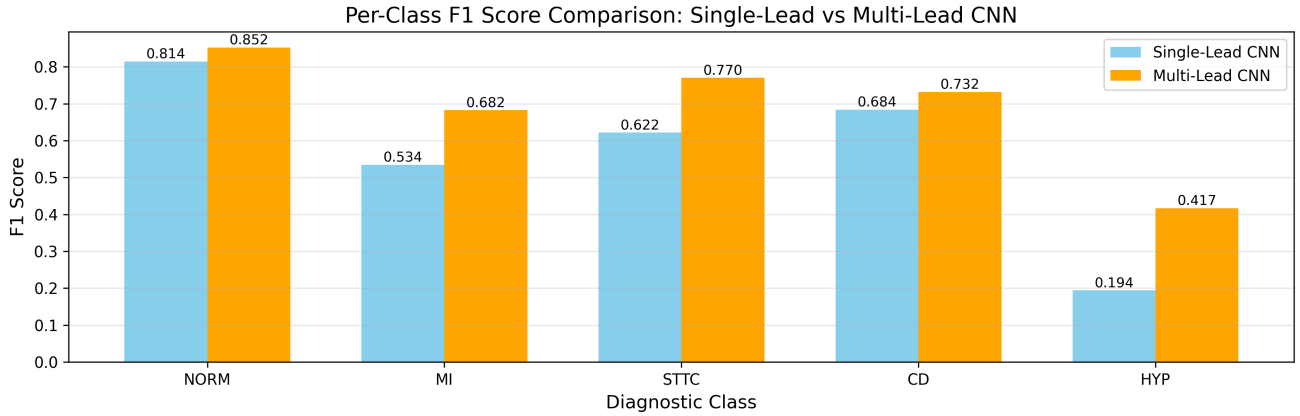

Fig. 5: Per-class F1-scores comparing Single-Lead and Multi-Lead CNNs. Multi-lead processing provides substantial gains across all classes, particularly for MI, STTC, and HYP which require spatial localization.

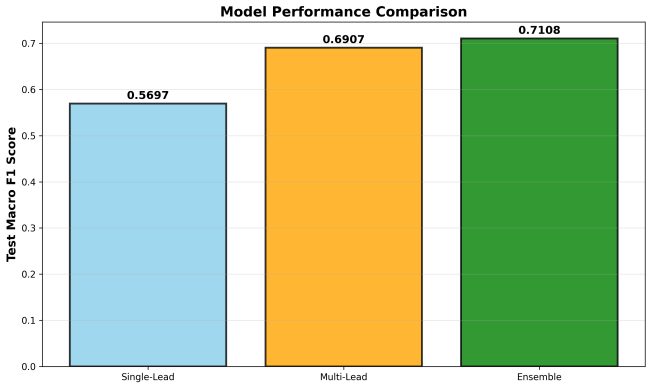

Fig. 6: Progressive performance improvement: Single-Lead (0.5697) → Multi-Lead (0.6907) → Ensemble (0.7108), demonstrating the value of multi-lead spatial information and ensemble learning.

on PTB-XL: Zhou et al. [5] achieved 0.73, Kang et al.[6] reached 0.75, and Chen et al. [9] attained 0.76. Although our method is a bit below the current state-of-the-art, it is still efficient and simple. Most importantly, our systematic comparison shows that spatial information from multi-lead accounts for 86% of the performance improvement, with ensemble learning contributing the remaining 14

V. Conclusion

On the PTB-XL dataset, this research has analyzed different deep learning methods with a great level of detail for multi-lead ECG classification. Single-lead, multi-lead, and ensemble CNN architectures were compared in our stepwise study, resulting in a total F1 measure improvement of 24.8

Three crucial points were discovered. (1) Multi-lead spatial information accounts for most of the accuracy improvement (86% of total), especially for spatially localized conditions: myocardial infarction got better by 27.7% and hypertrophy increased more than twice (+114.5%). (2) Homogeneous ensembles continued to increase stability (+2.9%) with the biggest improvements for hard classes like MI (+7.0%). (3) Our technique achieves recent state-of-the-art results (F1=0.70–0.76) while being straightforward and computationally efficient. Measuring the share of multi-lead fusion and ensembling in the overall improvement gives a lot of room to future ECG analysis systems, which can be very helpful in practical terms. The work ahead includes exploring larger or more advanced ensembles, devising strategies for rare classes, ensuring robustness to clinical noise, and incorporating patient metadata or temporal information. Clinical validation is, therefore, the essential next step for real-world deployment.

References

[1] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, no. 1, p. 154, 2020.

[2] T. Vu, N. S. Digumarthy, and S. Mishra, "Real-time arrhythmia detection using convolutional neural network," *Frontiers in Big Data*, vol. 6, p. 1270756, 2023.

[3] S. Afzal, M. Maqsood, T. Nazir, U. Khan, F. Aadil, K. M. Awan, I. Mehmood, and O.-Y. Song, "Automated detection of arrhythmia in ECG signals using convolutional neural networks," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 3, pp. 710–717, 2024.

[4] N. Mehta, M. Hariharan, N. Prasad *et al.*, "Review of ECG detection and classification based on deep learning: Coherent taxonomy, motivation, open challenges and recommendations," *Biomedical Signal Processing and Control*, vol. 74, p. 103493, 2022.

[5] Y. Zhou, J. Wang, Q. Li, and W. Zhang, "Classification of multi-lead ECG based on multiple scales and hierarchical features," *Scientific Reports*, vol. 15, no. 1, p. 8127, 2025.

[6] L. Kang, X. Fu, J. Vazquez-Corral, E. Valveny, and D. Karatzas, "Multi-label ECG classification via feature fusion with xLSTM," *arXiv preprint arXiv:2504.16101*, 2024, available: https://arxiv.org/abs/2504.16101.

[7] Z. Cai, C. Liu, M. Foschi, X. Wang, L. Li, Y. Xu, and M. Zhao, "Classification of multi-lead ECG with deep residual convolutional neural networks," in *2022 Computing in Cardiology (CinC)*, vol. 49. IEEE, 2022, pp. 1–4.

[8] S. Kumar, A. Mallik, and A. Kumar, "An improved method to detect arrhythmia using ensemble learning-based feature fusion method with ECG signals," *PLOS ONE*, vol. 19, no. 4, p. e0297551, 2024.

[9] J. Chen, T. Wang, Y. Zhang, Y. Chen, J. Hu, and J. Zhong, "Multi-expert ensemble ECG diagnostic algorithm using mutually hierarchical multi-label learning," *NPJ Biological Physics and Mechanisms*, vol. 1, no. 1, p. 10, 2024.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.