Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

HexagonalWarriorMamba: Superior Threshold-Dependent Multi-label Classification of 12-Lead ECG Cardiac Abnormalities

Huawei Jiang
Sungkyunkwan University
Husna Mutahira
Sogang University
Shibo Wei
Gwangju Institute of Science and Technology
Jianhang Li
Tianjin Normal University
Vladimir Shin
Financial University
Juneho Yi
Sungkyunkwan University
Dongryeol Ryu
Sungkyunkwan University
Wonyoung Park
Gwangju Institute of Science and Technology
Mannan Muhammad
mannan.muhammad@qmul.ac.uk
Queen Mary University of London

Article

Keywords:

Posted Date: May 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9418786/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

HexagonalWarriorMamba: Superior Threshold-Dependent Multi-label Classification of 12-Lead ECG Cardiac Abnormalities

Huawei Jiang[1], Husna Mutahira[2], Shibo Wei[3], Jiahang Li[4], Vladimir Shin[5], Juneho Yi[6], Dongryeol Ryu[3], Wonyoung Park[3,*], and Mannan Saeed Muhammad[7,*]

[1]Sungkyunkwan University, Department of Computer Science and Engineering, Suwon, 16419, South Korea
[2]Sogang University, Department of Computer Science and Engineering, Seoul, 04104, South Korea
[3]Gwangju Institute of Science and Technology, Department of Biomedical Science and Engineering, Gwangju, 61005, South Korea
[4]Tianjin Normal University, School of Artificial Intelligence, Tianjin, 300387, China
[5]Financial University under the Government of the Russian Federation, Department of Artificial Intelligence, Moscow, 125167, Russia
[6]Sungkyunkwan University, Department of Electrical and Computer Engineering, Suwon, 16419, South Korea
[7]Queen Mary University of London, School of Electronic Engineering and Computer Science, London, E14NS, United Kingdom
[*]corresponding author: parkwy0122@gist.ac.kr; mannan.muhammad@qmul.ac.uk

ABSTRACT

The accurate automated diagnosis of cardiac abnormalities from 12-lead electrocardiograms (ECGs) is critical for managing cardiovascular disease. However, detecting concurrent conditions remains a challenge for traditional deep learning models, which often have limited ability to model the long-range dependencies inherent in ECG signals. This manuscript proposes HexagonalWarriorMamba (HWMamba), a framework built on the Mamba architecture that processes 12-lead ECGs as single-channel 2D images rather than conventional 1D time series. By integrating a hierarchical architecture with a 2D Selective Scan mechanism, HWMamba is designed to model global context and complex spatial relationships within the data. The model is evaluated on the PhysioNet/Computing in Cardiology Challenge 2021 dataset, which includes 26 diagnostic labels and comprises recordings collected from seven institutions across four countries and three continents. Results demonstrate that HWMamba outperforms current state-of-the-art (SOTA) methods across five key threshold-dependent metrics, including Challenge Score and Subset Accuracy. These improvements provide a balance between strong discriminative capability and effective threshold selection derived from the training data, while maintaining near-SOTA performance in Macro AUROC. This Hexagonal Warrior performance, reflecting consistent performance across multiple evaluation dimensions, positions HWMamba as a robust and versatile approach for multi-label ECG classification.

Introduction

Electrocardiography (ECG) measures the electrical activity of the heart from multiple perspectives, enabling non-invasive assessment of cardiac function. As cardiovascular diseases remain a leading global cause of mortality, the need for timely and accurate diagnostic tools is well established[1]. Furthermore, the increasing adoption of wearable devices and remote monitoring systems has led to the generation of large volumes of ECG data, necessitating efficient and scalable processing approaches[2].

Large ECG datasets have made neural networks critical for detecting heart abnormalities[3–6]. Automated interpretation reduces the workload on healthcare professionals and standardizes diagnosis. This is particularly valuable in resource-limited settings lacking expert cardiologists. Conventional models such as Convolutional Neural Networks (CNNs) and Transformers have demonstrated strong performance in this domain[7,8], but may face limitations when modeling the long sequences characteristic of ECG signals[9]. More recently, selective State Space Models (SSMs), particularly Mamba[10], have been introduced as an alternative for handling long-range dependencies[11]. Initially developed for language modeling, Mamba has since been extended to 2D data through VMamba[12]. VMamba employs a Cross Scan Module to model global context with improved computational efficiency compared to Transformer-based approaches.

This paper presents HexagonalWarriorMamba (HWMamba). In contrast to conventional approaches that rely on 1D time-series inputs[13–16], the proposed method represents the 12-lead ECG as a single-channel image. This approach allows the use of established 2D image analysis architectures, including Swin Transformer[17], ConvNeXt[18], and VMamba, to model both

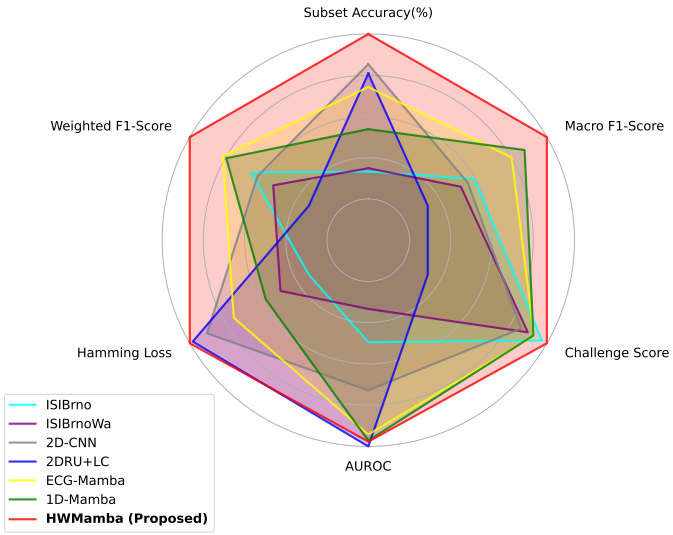


Figure 1. Performance Comparison of Algorithms (Hexagon Radar Chart)

inter- and intra-lead relationships within the 12-lead configuration. Accordingly, HWMamba adopts a hierarchical structure and incorporates a 2D Selective Scan (SS2D) mechanism to efficiently process 2D feature maps.

The proposed method was evaluated using publicly available datasets from the PhysioNet/Computing in Cardiology (CinC) Challenge 2021. The results indicate that HWMamba exhibits a balanced and consistent performance profile across multiple evaluation metrics. As illustrated in Figure 1, the model achieves state-of-the-art (SOTA) performance on five threshold-dependent metrics, including Subset Accuracy, Challenge Score, Hamming Loss, Macro F1-score, and Weighted F1-score, while maintaining near-SOTA performance on the threshold-independent Macro AUROC metric. The framework adopts a multi-label classification approach to better reflect clinical practice, where diagnoses often involve concurrent conditions[19,20], for example, Atrial Fibrillation occurring alongside Right Bundle Branch Block[21,22]. In contrast to single-label classification, which selects only the highest confidence prediction, multi-label settings require threshold-based evaluation to assess multiple conditions independently. Within this context, HWMamba demonstrates a balance between strong discriminative capability and effective threshold selection derived from the training data. This behavior is associated with improved performance relative to recent methods such as 2DRU+LC[23]. The consistent performance observed across all six evaluation dimensions supports the suitability of HWMamba as a robust approach for multi-label ECG classification.

The contributions of this paper are summarized as follows:

- **Benchmarking and Performance:** This study establishes a new reference point for multi-label ECG classification, showing that HWMamba outperforms current SOTA methods on several threshold-dependent metrics, including Challenge Score and Subset Accuracy. This “Hexagonal Warrior” performance reflects a balanced and consistent performance profile across evaluation dimensions, supporting robust identification of concurrent cardiac abnormalities relative to existing approaches.
- **Generalization and Threshold Stability:** The analysis of the performance gap between training and testing thresholds indicates that strong performance on threshold-independent metrics such as AUROC does not necessarily translate to practical effectiveness. This highlights the importance of stable threshold selection for reliable multi-label cardiac diagnosis in real-world settings.

Related Work

Recent advancements in the multi-label classification of 12-lead ECGs can be broadly grouped into two dominant processing paradigms[24]: treating the signal as a native 1D time series or transforming it into a structured 2D pseudo-image.

Classification on Native 1D Time-Series Input

The most conventional approach involves processing 12-lead ECG data via 1D CNNs, utilizing the temporal nature of the electrical signals. A representative example of this methodology is the solution proposed by team ISIBrno[13]. Their method secured first place in the PhysioNet/CinC Challenge 2021, achieving the highest Challenge Score. This method utilized an ensemble of residual networks enhanced with attention mechanisms to optimize feature extraction across leads. Subsequent

research by the same authors[14] revisited this architecture to examine the role of attention modules. In contrast to the common trend of increasing model complexity, their findings suggested that a simplified residual network without attention achieved slightly improved performance compared to the original attention-based design[13]. This observation highlights the importance of careful architectural design when optimizing 1D backbones for cardiac signal analysis.

More recently, the focus has shifted toward SSMs to capture long-range dependencies in physiological signals. Both ECG-Mamba[15] and 1D-Mamba[16] have adopted the Vision Mamba[25] architecture adapted for 1D signal processing. The promising results demonstrated by these methods on the AUROC metric for multi-label heart disease classification validate the efficacy of architectures based on Mamba for analyzing ECG time series.

Classification via Structured 2D Pseudo-Image Input

Parallel to 1D sequence modeling, a growing body of work explores the representation of 12-lead ECGs as rectangular 2D images to utilize established architectures[17,18] originally designed for computer vision. Within this paradigm, the 2D-CNN approach[26] represents the 12-lead ECG as a 2D data structure with a single channel. This method utilizes Deep Residual 2D CNNs and demonstrates the effectiveness of spatial convolution operations in modeling correlations between leads.

Among recent approaches, the 2DRU+LC method[23] reports strong performance on the threshold-independent AUROC metric. By adopting a U-Net architecture[27], originally developed for biomedical image segmentation, this method introduces a structured design for ECG classification and achieves competitive results on AUROC.

While existing 2D approaches demonstrate strong performance on threshold-independent metrics, the proposed HWMamba framework is designed to address limitations in clinical applicability. In particular, it shows improved performance on threshold-dependent metrics, which are important for real-world diagnostic settings, while maintaining competitive performance on AUROC with only a modest difference relative to the 2DRU+LC method.

Methods

Preliminaries

This study utilizes the capability of SSMs for modeling long-range dependencies in ECG signals. Originating from classical control theory and the Kalman filter[28], SSMs describe sequence modeling as a transformation of a 1D input signal $x(t) \in \mathbb{R}^{1\times 1}$ to an output signal $y(t) \in \mathbb{R}^{1\times 1}$ through a latent state representation $h(t) \in \mathbb{R}^{N\times 1}$.

State Space Models

The fundamental continuous-time SSM is defined by two equations: a state equation 1 and an output equation 2. Given an input signal $x(t)$, the system maps it to the latent state $h(t)$ and projects it to the output $y(t)$ as follows:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t) \tag{1}$$

$$y(t) = \mathbf{C}h(t) \tag{2}$$

Here, $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the state matrix, while $\mathbf{B} \in \mathbb{R}^{N\times 1}$ and $\mathbf{C} \in \mathbb{R}^{1\times N}$ serve as projection matrices. The $\mathbf{D}x(t)$ skip connection is excluded for simplicity.

Discretization

Processing ECG feature maps requires discretizing the continuous-time state space model. A timescale parameter Δ controls this transition by acting as the sampling step size. The Zero-Order Hold (ZOH) assumption treats the input as constant over each interval Δ and transforms the continuous parameters $(\mathbf{A}, \mathbf{B})$ into the discrete parameters $(\overline{\mathbf{A}}, \overline{\mathbf{B}})$ as follows:

$$\overline{\mathbf{A}} = \exp(\mathbf{\Delta A}) \tag{3}$$

$$\overline{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot \mathbf{\Delta B} \tag{4}$$

By applying discretization, the SSM is adapted for discrete sequence data at each time step k (see Supplementary Information for the complete mathematical proof).

$$h_k = \overline{\mathbf{A}}h_{k-1} + \overline{\mathbf{B}}x_k \tag{5}$$

$$y_k = \mathbf{C}h_k \tag{6}$$

Selective State Space Models (Mamba)

Mamba overcomes Linear Time-Invariant limitations through a selection mechanism computing the parameter set $\{\mathbf{B}, \mathbf{C}, \mathbf{\Delta}\}$ as functions of the input x_k rather than remaining static:

$$\mathbf{B}_k = \mathbf{W}_B x_k \tag{7}$$

$$\mathbf{C}_k = \mathbf{W}_C x_k \tag{8}$$

$$\mathbf{\Delta}_k = \text{softplus}(b_\Delta + \mathbf{W}_\Delta x_k) \tag{9}$$

where $\mathbf{W}_B, \mathbf{W}_C \in \mathbb{R}^{N \times D}$ and $\mathbf{W}_\Delta \in \mathbb{R}^{D \times D}$ are learnable projection matrices, with D denoting the input channel dimension and N the state dimension. The timescale projection $\mathbf{W}_\Delta$ is parameterized via a low-rank bottleneck of dimension R (typically $R \ll D$) to improve computational efficiency. A softplus activation is then applied to enforce positivity, ensuring that the generated timescale $\mathbf{\Delta}$ represents a mathematically valid, nonnegative sampling interval. Additionally, b_Δ is a learnable bias parameter controlling the default timescale. Unlike $\mathbf{\Delta}$, the parameters $\mathbf{B}$ and $\mathbf{C}$ are generated via bias-free projections.

ECG Sample Pre-Processing

All ECG signals were first uniformly resampled to a target rate of 500 Hz following established methodologies [13–16,29]. Signals originally at 1000 Hz were downsampled using polyphase filtering to minimize aliasing. Conversely, 257 Hz signals were upsampled to 500 Hz using the Fast Fourier Transform. Signal length was standardized by right zero-padding shorter records to 8192 samples, while longer records were randomly sampled and truncated to this uniform length.

Normalization was intentionally omitted for the 12-lead ECG signals, following[15] and in contrast to several commonly adopted practices[13,14,16,23,26]. This decision preserves diagnostic information related to amplitude, which may otherwise be altered by standard normalization techniques such as Z-score scaling.

HexagonalWarriorMamba

The overall architecture of HWMamba is shown in Figure 2(a). Its hierarchical design is similar to that of the Swin Transformer[17] and VMamba[12], processing the input 12-lead ECG at multiple resolutions. This hierarchy is achieved through progressive downsampling, which transitions the representation from high resolution with shallow feature depth to lower resolution with deeper feature depth. The Swin Transformer and VMamba operate on square image inputs such as 224×224. In contrast, the proposed model processes the 12-lead ECG as a rectangular input, characterized by a short spatial dimension of 12 leads and a long temporal dimension of 8192 samples. This difference motivates a modification to the initial convolutional layer of HWMamba. The first Conv2d layer utilizes a rectangular kernel of size (3, 16) with a stride of (1, 16) based on insights from 2D-CNN and 2DRU+LC to better accommodate the asymmetric structure of the ECG signals.

The model then applies multiple HWMamba blocks and downsampling layers. The number of blocks at each stage is empirically determined as [2, 2, 2, 8], with corresponding feature dimensions of [48, 96, 192, 384], as shown in Figure 2(a). A final classifier layer completes the architecture as detailed in Figure 2(c). This layer incorporates layer normalization to stabilize feature distributions, followed by global average pooling and flattening to aggregate spatial information before a fully connected layer produces the final multi-label classification outputs.

In Figure 2(b), the SS2D module and the Depthwise Conv2d[30] are positioned between two linear layers. The SS2D module is a core component of VMamba. As illustrated in Figure 3, neither Vision Mamba nor VMamba fundamentally alters the underlying Mamba mechanism for processing sequence data; rather, they differ in how they serialize spatial data. The standard Mamba operation on an input sequence $\mathbf{X}$ is defined as follows, where $\mathbf{Y}$ represents the output sequence:

$$\mathbf{Y} = \text{S6}(\mathbf{X}) \tag{10}$$

Vision Mamba extends this by flattening the feature map and augmenting it with a flipped sequence to scan forward and backward, as expressed by:

$$\mathbf{Y}_{\text{forward}} = \text{S6}(\mathbf{X}_{\text{forward}}) \tag{11}$$

$$\mathbf{Y}_{\text{backward}} = \text{S6}(\mathbf{X}_{\text{backward}}) \tag{12}$$

In contrast, VMamba transforms the 2D spatial structure into 1D sequences by scanning in four distinct directions (cross-scan). This process is formulated as:

$$\mathbf{Y}_{\text{left}\to\text{right}} = \text{S6}(\mathbf{X}_{\text{left}\to\text{right}}) \tag{13}$$

$$\mathbf{Y}_{\text{top}\to\text{bottom}} = \text{S6}(\mathbf{X}_{\text{top}\to\text{bottom}}) \tag{14}$$

$$\mathbf{Y}_{\text{right}\to\text{left}} = \text{S6}(\mathbf{X}_{\text{right}\to\text{left}}) \tag{15}$$

$$\mathbf{Y}_{\text{bottom}\to\text{top}} = \text{S6}(\mathbf{X}_{\text{bottom}\to\text{top}}) \tag{16}$$

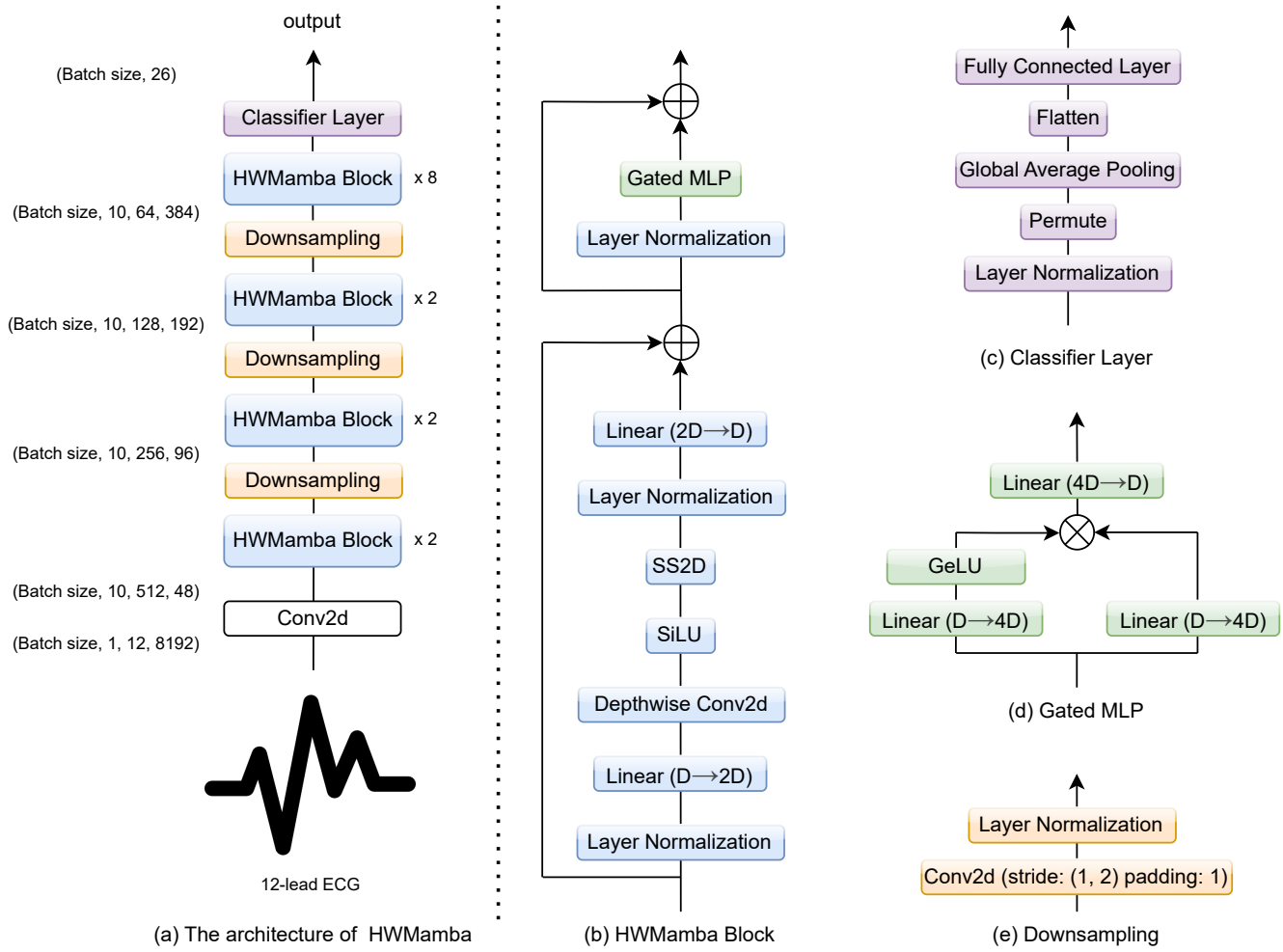


Figure 2. HWMamba architecture and submodules. (a) Full model framework; (b) HWMamba block; (c) Classifier layer; (d) Gated MLP; and (e) Downsampling module.

An SS2D approach based on VMamba is adopted to model spatial dependencies, as the proposed method represents the 12-lead ECG as an image to generate a 2D feature map. Depthwise Conv2d is employed to reduce computational cost by decomposing standard convolution into two efficient steps. A depthwise convolution first applies one spatial filter per channel independently before a pointwise 1×1 convolution mixes information across channels. This significantly reduces the number of parameters and computational complexity compared to standard convolutions.

The two linear layers surrounding the SS2D and depthwise convolution form an expansion-contraction MLP block similar to the FFN in Transformers and the MLP block in Vision Mamba/VMamba. This block expands the feature dimension before projecting it back to the original size. As illustrated in Figure 2(b), the channel dimension is expanded by a factor of 2 before reduction. In contrast, the Gated MLP[31,32] shown in Figure 2(d) expands the feature dimension by a factor of 4 prior to projection. The gating mechanism introduces a multiplicative interaction between features and represents the key reason for using a Gated MLP instead of a standard MLP. It enables adaptive feature selection and contributes to more stable training while preserving informative representations.

Figure 2(e) illustrates the downsampling module, which consists of a Conv2d layer followed by Layer Normalization. The Conv2d uses a 3×3 kernel. The stride is empirically set to $(1,2)$ to account for the feature map shown in Figure 2(a). A padding of 1 is applied following the standard downsampling configuration utilized in VMamba. The output is subsequently normalized to the range $[0,1]$ for evaluation using threshold-dependent metrics.

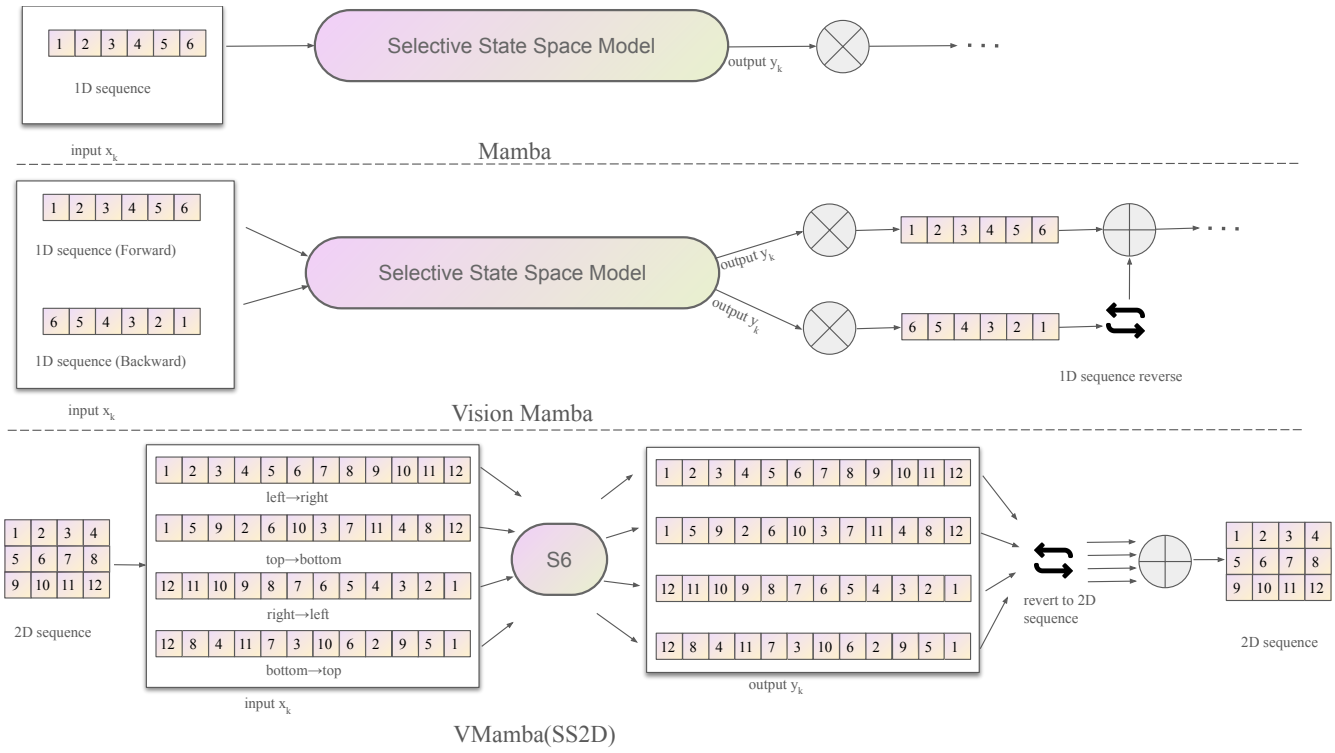


Figure 3. Comparison of data sequencing strategies for Selective State Space Models(S6) in Mamba, Vision Mamba and VMamba architectures

Dataset Description

The experimental data used in this study are sourced from the PhysioNet/CinC Challenge 2021[33,34] and consist of 88,253 multi-lead ECGs from seven international institutions. These recordings cover 26 heart conditions (25 abnormalities plus sinus rhythm) and provide multi-label annotations for each. The signals utilize diverse sampling frequencies including 500 Hz (n=87,663) alongside 1000 Hz (n=516) and 257 Hz (n=74). Table 1 details the specific subset utilized in this study.

Experiments

Experimental Setup

Both the training and evaluation procedures were implemented using the PyTorch deep learning framework. All experiments were conducted on a GeForce RTX 3090 Ti GPU with 24 GB of dedicated memory, running Ubuntu 22.04 with CUDA Toolkit 11.8 (Driver 535).

The Adam[35] optimizer was used with a base learning rate of 0.001 and a batch size of 20. The optimizer parameters were

Table 1. Summary of the ECG datasets sourced from multiple institutions via the PhysioNet/CinC Challenge 2021 utilized in this study.

Source Database	Recordings (*N*)	Duration (s)	Frequency (Hz)
CPSC / CPSC-Extra	10,330	6–144	500
St Petersburg INCART	74	1,800	257
PTB / PTB-XL	22,353	10–120	500 / 1000
Georgia (G12EC)	10,344	5–10	500
Ningbo First Hospital	34,905	10	500
Chapman / Shaoxing	10,247	10	500
Total Aggregate	**88,253**	—	—

set to $\beta_1 = 0.9$ alongside $\beta_2 = 0.98$ and $\varepsilon = 1 \times 10^{-9}$. A cosine annealing scheduler[36] with a warmup phase of 5 epochs was employed to manage the learning rate and align with standard practices in Mamba[10,12,25] and Transformer[17,37] architectures. This warmup phase linearly increased the learning rate from an initial 1×10^{-5} to the base 0.001. This was followed by cosine decay over 13 epochs, reaching a minimum learning rate of 1×10^{-6}. The choice of optimizer and scheduler configuration was determined empirically.

The dataset was split into training and testing subsets using stratified proportional allocation and random shuffling. This technique was essential due to the 26 imbalanced classes. The overall data split designated 80% for training and the remaining 20% for testing. A 5-fold cross-validation strategy was adopted in place of a separate validation set to provide a more robust evaluation of model performance while maximizing the use of available data. This approach is particularly beneficial for imbalanced datasets. The same training and testing splits were consistently used across experiments to enable direct comparison with SOTA methods. Each fold consisted of 70,602 training records and 17,651 testing records.

A systematic threshold search strategy was applied during the validation phase to evaluate the threshold-dependent performance metrics. The model outputs a probability vector $\hat{y} \in [0,1]^C$ where $C = 26$ denotes the number of cardiac conditions. A global decision threshold τ is used to convert these continuous probabilities into binary predictions. Following standard practice in multi-label classification, a grid search over $\tau \in \{0, 0.02, \ldots, 1\}$ was performed. The binary prediction for the i-th class, denoted as $\hat{y}_{i,\text{binary}}$, for a given threshold τ, is defined as:

$$\hat{y}_{i,\text{binary}} = \begin{cases} 1 & \text{if } \hat{y}_i > \tau \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Evaluation Metrics

Six evaluation metrics were employed to compare the proposed HWMamba with SOTA models: Subset Accuracy, Challenge Score, Hamming Loss, Macro F1-score, Weighted F1-score, and Macro AUROC. Among these, AUROC is the only threshold-independent metric, while the remaining five are threshold-dependent.

Individual metrics such as Sensitivity, Specificity, and Precision are not considered primary indicators in this study. These metrics can attain extreme values, for example 1.0, under trivial threshold settings (e.g., $\tau = 0$ or $\tau = 1$), and may therefore not provide a balanced evaluation. Macro and Weighted F1-scores are adopted instead because they require a trade-off between precision and recall.

Furthermore, AUROC is selected over AUPRC to provide a comprehensive assessment of diagnostic performance. While AUPRC focuses exclusively on the positive class (ignoring true negatives), AUROC incorporates the false positive rate, enabling evaluation of the model's ability to distinguish between positive classes and the negative background. Since the precision-recall trade-off is already reflected by the F1-scores, AUROC serves as a complementary metric for assessing overall ranking performance across both sensitivity and specificity.

Subset Accuracy

Accuracy is commonly used to evaluate the performance of single-label classification tasks. However, in contrast to single-label settings, ECG diagnosis often involves multiple diagnostic labels [19], making it a multi-label classification problem. In this context, subset accuracy provides a more appropriate evaluation metric, as defined below:

$$Acc_{\text{subset}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(Y_i = Z_i) \tag{18}$$

where Acc_{subset} denotes the subset accuracy metric, N is the total number of samples, and i is the sample index. Y_i represents the true set of labels for sample i, while Z_i denotes the predicted set of labels for sample i. The indicator function $\mathbb{I}(\cdot)$ equals 1 when $Y_i = Z_i$ (i.e., true match), and 0 otherwise.

Challenge Score

The Challenge Score is a specialized evaluation metric introduced for the PhysioNet/Computing in Cardiology (CinC) 2021 Challenge. It extends the scoring framework used in the 2020 Challenge[38] by incorporating a broader range of diagnostic classes and clinically informed weighting. The metric is computed using a predefined weight matrix (e.g., `weights.csv`), which encodes the clinical relationships between different cardiac conditions.

Unlike standard accuracy, the Challenge Score accounts for the clinical significance of predictions. It assigns partial credit to misclassifications when the predicted condition is clinically similar to the true diagnosis, while applying stronger penalties to errors that may have greater clinical impact. This weighting strategy enables a more clinically meaningful evaluation by reflecting the relative importance of different types of diagnostic errors. Ultimately, incorporating the Challenge Score metric can bridge deep learning algorithms with clinical realities.

Table 2. Comparison of different HWMamba configurations for multi-label ECG classification. The analysis highlights the impact of classifier architecture (Standard vs. Gated MLP), stride patterns, normalization techniques, and state dimensions. Best results are **bolded**, and second-best results are underlined.

Configuration	Subset Accuracy(%)	Challenge score	Hamming ↓	$F1_{Macro}$	$F1_{Weighted}$	AUROC
HWMamba (MLP)	**60.60**	0.7106	**0.0240**	0.5951	0.7603	0.9686
HWMamba (Stride: 2,2)	58.08	0.7079	0.0266	0.5979	0.7593	0.9672
HWMamba (Stride: 1,3)	57.37	0.6937	0.0262	0.5682	0.7348	0.9626
HWMamba (w/ Norm)	59.06	0.7138	0.0247	0.5982	0.7640	0.9655
HWMamba (State Dim: 1)	59.00	0.7215	0.0252	0.5999	0.7698	**0.9703**
HWMamba (Proposed)	59.51	**0.7294**	0.0250	**0.6156**	**0.7730**	0.9698

Hamming Loss

The Hamming Loss (Hamming) is an instance-based metric for evaluating multi-label classification models, which measures the fraction of individual labels that are incorrectly predicted. It is computed by comparing the predicted label set with the ground truth set across all samples, counting the total number of label errors and normalizing this count by the total number of labels over all instances. Label-level errors occur when a label is incorrectly predicted as present or absent. The formulation is given below:

$$\text{Hamming} = \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{j=1}^{L} \mathbb{I}(\hat{y}_{i,j} \neq y_{i,j}) \tag{19}$$

N is the total number of samples. L is the total number of labels. $y_{i,j} \in \{0,1\}$ is the true label for sample i and label j. $\hat{y}_{i,j} \in \{0,1\}$ is the predicted label for sample i and label j. $\mathbb{I}(\cdot)$ is the Indicator Function, which returns 1 if the condition is true (i.e., a misclassification) and 0 otherwise.

Macro F1-score and Weighted F1-score

The F1-score for an individual label i is the harmonic mean of its Precision (P_i) and Recall (R_i):

$$F1_i = \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \tag{20}$$

The Macro F1-score ($F1_{Macro}$) is calculated as the unweighted arithmetic mean of the individual F1-scores ($F1_i$) across all N labels, treating each label equally:

$$F1_{Macro} = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{21}$$

The Weighted F1-score ($F1_{Weighted}$) averages the individual F1-scores ($F1_i$), where each is weighted by the support (S_i), which is the number of true instances for label i.

$$F1_{Weighted} = \sum_{i=1}^{N} \left(F1_i \cdot \frac{S_i}{M} \right) \tag{22}$$

where S_i is the support (number of true instances) for label i, and M is the total number of samples, $M = \sum_{i=1}^{N} S_i$.

AUROC

Macro AUROC is a threshold-independent evaluation metric that assesses a classifier's ability to distinguish between positive and negative classes. It is computed by averaging the area under the receiver operating characteristic curve across all classes, assigning equal importance to each class regardless of class frequency. This ensures that the model's overall discriminative capability is evaluated in a balanced manner, even in the presence of class imbalance.

Results and Discussion

Table 2 summarizes the performance of various HWMamba configurations, providing insights into the optimal architectural and preprocessing choices for 12-lead ECG analysis. Unlike the standard VMamba architecture, which is typically designed for 2D image inputs, the experimental results indicate that task-specific adaptations are required for multi-label ECG classification.

Table 3. Comparison of HWMamba (proposed) with SOTA Methods on the PhysioNet/CinC Challenge 2021 Dataset. Best results are **bolded**, and second-best results are underlined.

Method	Subset Accuracy(%)	Challenge score	Hamming ↓	$F1_{Macro}$	$F1_{Weighted}$	AUROC
ISIBrno	50.19	0.7187	0.0365	0.5281	0.7282	0.9048
ISIBrnoWa	50.42	0.6873	0.0329	0.5119	0.7120	0.8831
2D-CNN	57.70	0.6726	0.0262	0.5202	0.7235	0.9363
2DRU+LC	56.67	0.4659	0.0252	0.4720	0.6860	**0.9729**
ECG-Mamba	56.10	0.7010	0.0283	0.5731	0.7488	0.9651
1D-Mamba	53.15	0.6996	0.0313	0.5887	0.7463	0.9691
HWMamba (Proposed)	**59.51**	**0.7294**	**0.0250**	**0.6156**	**0.7730**	0.9698

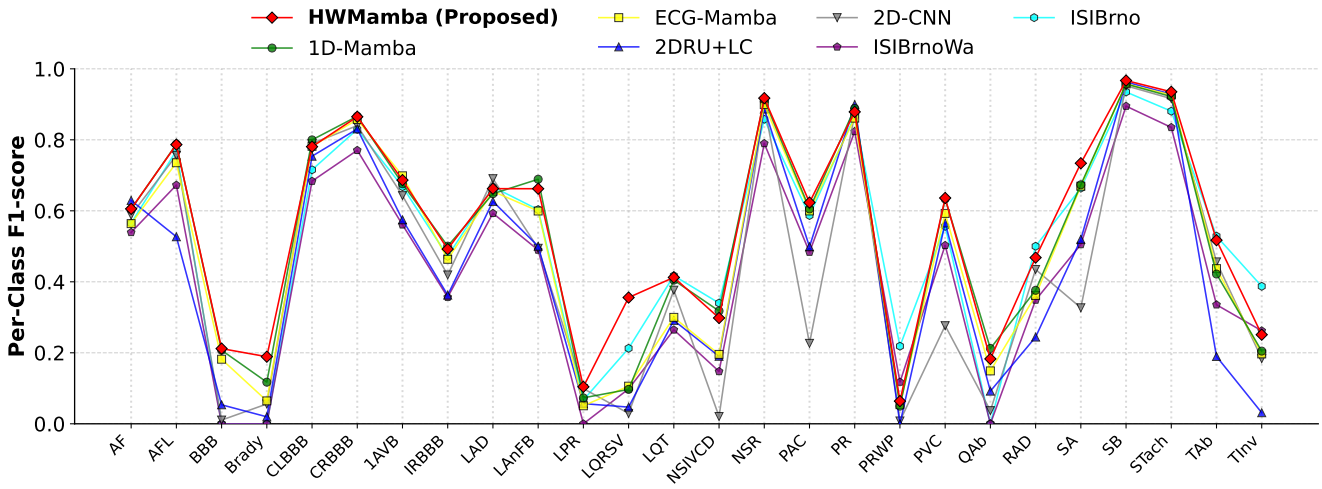


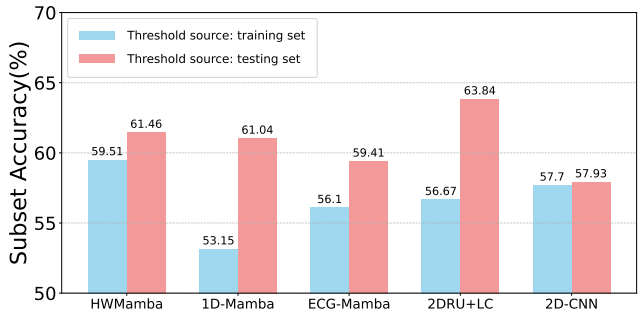
Figure 4. Per-Class F1-Score Comparison of HWMamba and SOTA Methods

The model achieves peak performance with a downsampling stride of (1, 2) and a state dimension of 16. The (1, 2) stride effectively reduces the temporal dimension while preserving critical inter-lead spatial information, outperforming both symmetric (2, 2) and alternative asymmetric (1, 3) strides. Furthermore, utilizing the Gated MLP proved essential for generalization. While a standard MLP yielded higher subset accuracy, the Gated MLP configuration (the final HWMamba) achieved the highest Challenge Score (0.7294) and F1 Weighted (0.7730).

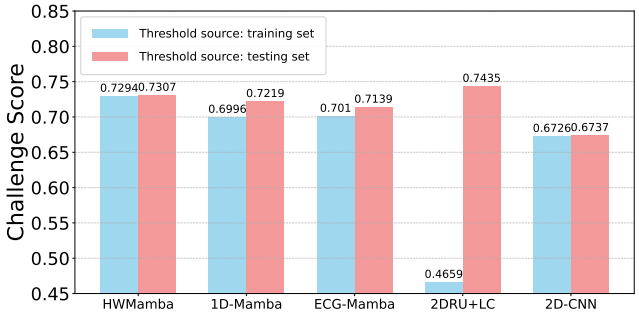
A notable observation relates to the role of input normalization. Although normalization is widely adopted in existing literature as a standard preprocessing step, its application was observed to be suboptimal for this specific task. As observed in the "HWMamba (w/ Norm)" configuration, applying Z-score normalization results in a decrease across most evaluation metrics, with the Challenge Score reduced to 0.7138 and the Macro F1-score to 0.5982 compared to the unnormalized baseline. This suggests that amplitude-related characteristics and the original signal distribution contain discriminative information that is important for the HWMamba model, and may be altered by standard normalization techniques.

Table 3 presents a comprehensive quantitative comparison between the proposed HWMamba and current SOTA methods on the PhysioNet/CinC Challenge 2021 dataset. The proposed model was evaluated against three distinct categories of baselines: (1) leading methods from the original challenge, including the winner, ISIBrno, and its updated variant ISIBrnoWa; (2) deep learning approaches based on images, such as 2D CNN and 2DRU+LC, which treat 12-lead ECGs as images with a single channel; and (3) recent Mamba architectures, ECG-Mamba and 1D-Mamba.

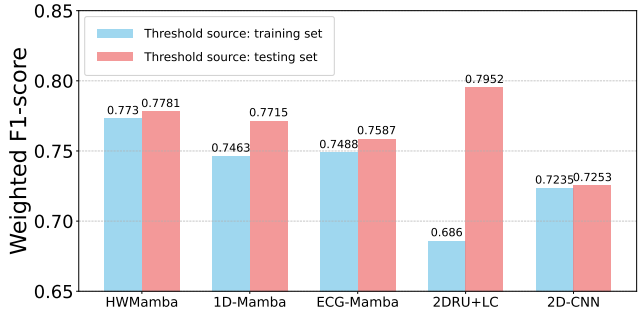
To adhere to the principle of no data leakage, the decision thresholds for all threshold-dependent metrics were optimized exclusively on the training set. As shown in Table 3, HWMamba demonstrates improved performance across threshold-dependent metrics. In particular, the model achieves a Challenge Score of 0.7294, exceeding that of the previous challenge winner ISIBrno (0.7187). HWMamba also shows consistent performance across complementary metrics, achieving the highest Subset Accuracy (59.51%), Macro F1-score (0.6156), and Weighted F1-score (0.7730), while maintaining the lowest Hamming Loss (0.0250). These results suggest that the model provides a balanced performance profile across multiple evaluation dimensions.



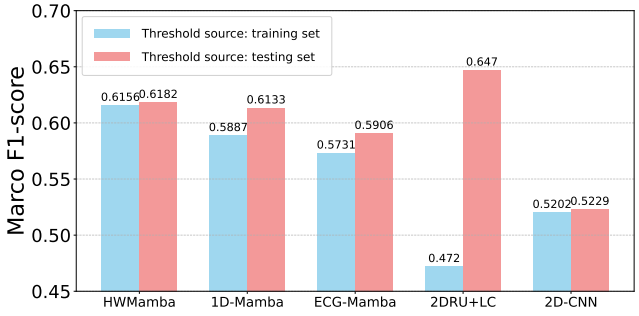
(a) Subset Accuracy Comparison: Training vs. Testing Thresholds



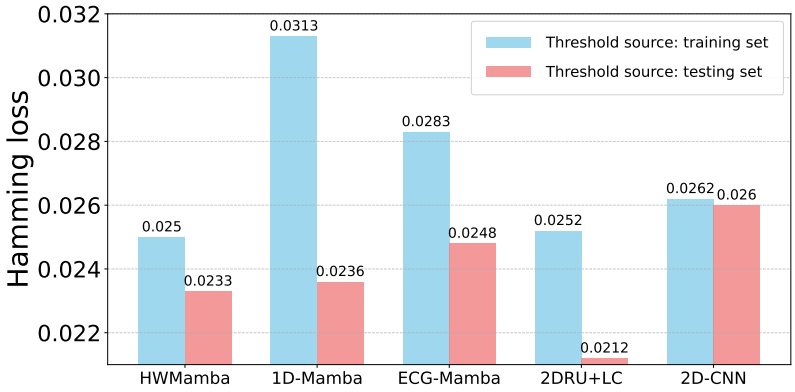
(b) Challenge Score Comparison: Training vs. Testing Thresholds



(c) Weighted F1-score Comparison: Training vs. Testing Thresholds



(d) Macro F1-score Comparison: Training vs. Testing Thresholds



(e) Hamming Loss Comparison: Training vs. Testing Thresholds(lower is better)

Figure 5. Impact of Threshold Source on Various Performance Metrics Across Five Models

Figure 4 illustrates the per-class F1-score comparison between HWMamba and SOTA models. The F1-score is chosen because Subset Accuracy, Challenge Score, and Hamming Loss are evaluated at the instance level and cannot assess individual classes. Threshold-independent metrics like AUROC are also excluded to maintain focus on threshold-dependent evaluation. (Abbreviations for diagnosis labels can be found in the Supplementary Information.) HWMamba consistently outperforms baselines across most cardiac abnormality classes. It achieves F1-scores approaching or exceeding 0.8 in categories like NSR, SB, STach, and CRBBB. Furthermore, HWMamba demonstrates notable resilience in challenging classes such as LPR and LQRSV, where other models struggle. This consistent performance highlights the robustness of the proposed architecture for multi-label ECG classification.

In terms of threshold-independent metrics, the U-Net-adapted method 2DRU+LC remains the SOTA for AUROC (0.9729). However, HWMamba remains highly competitive with an AUROC of 0.9698, trailing by a narrow margin. While 2DRU+LC achieves a marginally higher AUROC, HWMamba significantly outperforms it on all threshold-dependent metrics (Accuracy, F1, and Challenge Score), demonstrating superior capability in making definitive diagnostic predictions.

To examine the distribution characteristics of the HWMamba across the training and testing sets, Figure 5 presents a comparative analysis of five models across five threshold-dependent performance metrics: Figure 5a Subset Accuracy, Figure 5b Challenge Score, Figure 5c Weighted F1-score, Figure 5d Macro F1-score, and Figure 5e Hamming Loss (where lower

values indicate better performance). The ISIBrno model and its upgraded variant are excluded from this analysis. This is because Team ISIBrno uses a custom sparsity loss that constrains the outputs to binary values (0 or 1), resulting in an optimal decision threshold that remains fixed near 1. Consequently, it is not possible to meaningfully analyze threshold-dependent distributional differences between the training and testing sets for this model, unlike those trained using Binary Cross-Entropy.

The bar charts illustrate the impact of the data source used to determine the optimal decision thresholds. The blue bars represent performance when thresholds are calibrated on the training set (realistic scenario), while red bars indicate thresholds optimized on the testing set, representing a theoretical upper bound on performance. The gap between the training and testing thresholds reflects the models' robustness to variations in data distribution between the subsets. A larger disparity indicates decreased robustness and a higher sensitivity to domain shift between training and testing data.

In terms of generalization performance, the 2D-CNN model exhibits the smallest gap between threshold sources across most metrics. Although it does not achieve the highest overall scores, its performance remains stable, indicating strong generalization capabilities and effective handling of distributional consistency between datasets.

In contrast, the 2DRU+LC model demonstrates the largest performance gap, particularly evident in the Challenge Score (Figure 5b) and Hamming Loss (Figure 5e). While performance appears strong using thresholds optimized on test data, the substantially lower results using thresholds derived from training data indicate poor generalization likely caused by sensitivity to distributional mismatches.

The proposed HWMamba method achieves a favorable balance between performance and robustness. It consistently delivers competitive results across all evaluated metrics while maintaining a relatively small gap between threshold settings derived from training and testing. This indicates that the model not only achieves strong predictive performance but also demonstrates improved resilience to distributional variation between datasets.

Conclusions

This study introduced the HWMamba framework to redefine the automated diagnosis of cardiac abnormalities by processing 12-lead ECG signals as 2D representations with a single channel. Integrating a hierarchical architecture with the SS2D mechanism allows HWMamba to successfully overcome the limitations of conventional 1D time series models in capturing dependencies over long ranges and complex spatial relationships within ECG data. The comprehensive evaluation on the large-scale PhysioNet/CinC Challenge 2021 dataset demonstrates that HWMamba establishes a benchmark for multi-label ECG classification. The model achieved SOTA performance across five critical threshold-dependent metrics, including Subset Accuracy, Challenge Score, and Weighted F1-score while maintaining near-SOTA performance in AUROC. This balanced performance profile highlights the model's ability to simultaneously optimize multiple evaluation criteria, indicating both high predictive accuracy and robustness in multi-label cardiac diagnosis.

Further analysis reveals that standard input normalization leads to a measurable degradation in performance, suggesting that absolute ECG amplitude carries diagnostically relevant information that is effectively preserved and exploited by HWMamba. In addition, the relatively small performance gap between training-derived and testing-derived decision thresholds indicates strong generalization capability, which is critical for reliable deployment in clinical settings. In contrast, the 2DRU+LC approach based on the U-Net architecture exhibits limited generalization when thresholds are derived from the training set, despite achieving strong performance under optimally tuned thresholds. This observation underscores the sensitivity of certain architectures to distributional shifts and highlights an important direction for future research. In particular, combining the structural advantages of U-Net with the sequence modeling capabilities of Mamba-based architectures may provide a promising pathway to improve both threshold robustness and overall classification performance.

Data availability

The datasets analyzed during the current study were sourced from the Official PhysioNet/CinC Challenge 2021. The data is publicly available in the PhysioNet repository and can be accessed at the following stable URL: https://physionet.org/content/challenge-2021/1.0.3/.

References

1. Roth, G. A. *et al.* Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J. Am. college cardiology* **70**, 1–25 (2017).
2. Prieto-Avalos, G. *et al.* Wearable devices for physical monitoring of heart: a review. *Biosensors* **12**, 292 (2022).
3. Ribeiro, A. H. *et al.* Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nat. communications* **11**, 1760 (2020).

4. Siontis, K. C., Noseworthy, P. A., Attia, Z. I. & Friedman, P. A. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat. Rev. Cardiol.* **18**, 465–478 (2021).
5. Abrar, M. F. *et al.* Automated heart disease detection using swin transformer and ecg signal processing: a high-accuracy approach. *Sci. Reports* **15**, 39338 (2025).
6. Dhandapani, S., Somasundaram, H. & Angamuthu, T. Hybrid deep learning framework for heart disease prediction using ecg signal images. *Sci. Reports* **15**, 33922 (2025).
7. Zhao, Y., Cheng, J., Zhang, P. & Peng, X. Ecg classification using deep cnn improved by wavelet transform. *Comput. Mater. Continua* (2020).
8. Che, C., Zhang, P., Zhu, M., Qu, Y. & Jin, B. Constrained transformer network for ecg signal processing and arrhythmia classification. *BMC Med. Informatics Decis. Mak.* **21**, 184 (2021).
9. Ansari, M. Y. *et al.* A survey of transformers and large language models for ecg diagnosis: advances, challenges, and future directions. *Artif. Intell. Rev.* **58**, 261 (2025).
10. Gu, A. & Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling* (2024).
11. Zhang, Z. *et al.* Switch-umamba: Dynamic scanning vision mamba unet for medical image segmentation. *Med. Image Analysis* 103792 (2025).
12. Liu, Y. *et al.* Vmamba: Visual state space model. *Adv. neural information processing systems* **37**, 103031–103063 (2024).
13. Nejedly, P. *et al.* Classification of ecg using ensemble of residual cnns with attention mechanism. In *2021 Computing in Cardiology (CinC)*, vol. 48, 1–4 (IEEE, 2021).
14. Nejedly, P. *et al.* Classification of ecg using ensemble of residual cnns with or without attention mechanism. *Physiol. Meas.* **43**, 044001 (2022).
15. Jiang, H., Mutahira, H., Wei, S. & Muhammad, M. S. Ecg-mamba: Cardiac abnormality classification with non-uniform-mix augmentation on 12-lead ecgs. *IEEE J. Transl. Eng. Heal. Medicine* **13**, 461–470, DOI: 10.1109/JTEHM.2025.3613609 (2025).
16. Jiang, H., Mutahira, H., Huang, G. & Muhammad, M. S. One dimensional cnn ecg mamba for multilabel abnormality classification in 12 lead ecg (2025). 2510.13046.
17. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).
18. Liu, Z. *et al.* A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986 (2022).
19. Li, J. *et al.* An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI* **2**, AIoa2401033 (2025).
20. Sun, L., Li, C., Ren, Y. & Zhang, Y. A multitask dynamic graph attention autoencoder for imbalanced multilabel time series classification. *IEEE Transactions on Neural Networks Learn. Syst.* **35**, 11829–11842 (2024).
21. Zhang, F.-T. *et al.* Association between complete right bundle branch block and atrial fibrillation development. *Annals Noninvasive Electrocardiol.* **27**, e12966 (2022).
22. Ran, S. *et al.* Label correlation embedding guided network for multi-label ecg arrhythmia diagnosis. *Knowledge-Based Syst.* **270**, 110545 (2023).
23. Hwang, S., Cha, J., Heo, J., Cho, S. & Park, Y. Multi-label abnormality classification from 12-lead ecg using a 2d residual u-net. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2265–2269 (IEEE, 2024).
24. Le, D., Truong, S., Brijesh, P., Adjeroh, D. A. & Le, N. scl-st: Supervised contrastive learning with semantic transformations for multiple lead ecg arrhythmia classification. *IEEE journal biomedical health informatics* **27**, 2818–2828 (2023).
25. Zhu, L. *et al.* Vision mamba: efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, 62429–62442 (2024).
26. Elyamani, H. A., Salem, M. A., Melgani, F. & Yhiea, N. Deep residual 2d convolutional neural network for cardiovascular disease classification. *Sci. Reports* **14**, 22040 (2024).
27. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).

28. Kalman, R. E. *et al.* A new approach to linear filtering and prediction problems [j]. *J. basic Eng.* **82**, 35–45 (1960).

29. Natarajan, A. *et al.* A wide and deep transformer neural network for 12-lead ecg classification. In *2020 Computing in Cardiology*, 1–4 (IEEE, 2020).

30. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258 (2017).

31. Liu, H., Dai, Z., So, D. & Le, Q. V. Pay attention to mlps. *Adv. neural information processing systems* **34**, 9204–9215 (2021).

32. De, S. *et al.* Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427* (2024).

33. Reyna, M. A. *et al.* Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. PhysioNet. Available from: https://physionet.org/content/challenge-2021/1.0.3/ (2021).

34. Reyna, M. A. *et al.* Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021. In *2021 computing in cardiology (CinC)*, vol. 48, 1–4 (IEEE, 2021).

35. Kinga, D., Adam, J. B. *et al.* A method for stochastic optimization. In *International conference on learning representations (ICLR)*, vol. 5 (California;, 2015).

36. Loshchilov, I. & Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations* (2017).

37. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

38. Alday, E. A. P. *et al.* Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *Physiol. measurement* **41**, 124003 (2020).

Author contributions statement

H.J. conceived the original idea, designed and performed the simulations, and wrote the main manuscript. V.S. and M.S.M. developed and finalized the mathematical modeling for the State Space Model. D.R., W.P., and S.W. interpreted the 12-lead ECGs for heart disease detection. J.L. prepared the figures and graphical representations. H.M. and J.Y. reviewed and edited the manuscript, significantly improving its technical clarity. All authors reviewed and approved the final manuscript.

Funding

This study was supported by the National Research Foundation of Korea (NRF), funded by the Korean government (Ministry of Science and ICT) (RS-2025-00519038).

Additional information

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HWMambaScientificReportsSupplementaryInformation.pdf