

Vol 2 Issue 03 (2026)

<http://fsrma.org.hostajournal.com/>

ONLINE ISSN: 3134-7517

PRINT ISSN: 3134-7509

---

# A Lightweight Frequency-Domain Defense Framework for Adversarially Robust Wearable ECG Classification

**Kehang Lai**

School of Computing and Information Systems

The University of Melbourne

Melbourne, Victoria, Australia

Email: [cormaclai@gmail.com](mailto:cormaclai@gmail.com)

## Abstract.

*With the development of different kinds of wearable devices, more and more people are using these systems monitoring their body signals. In addition, deep neural networks show strong performance when it comes to single-lead ECG rhythm classification. However, those machine learning classifiers can be easily affected by small deliberately designed adversarial perturbations to original input data. Hence, this essay proposes a lightweight frequency-domain defense framework for protecting a single-lead ECG classifier against adversarial perturbations using the MIT-BIH Arrhythmia Database. The experiment in this essay uses both patient-level train/test separation and continuous multi-beat ECG windows, so the evaluation is closer to a real-world clinical setting where new patients are unseen during training. This framework uses FGSM and PGD to attack an unprotected baseline model trained by time-series based 1D CNN (one-dimensional convolutional neural network) and a proposed frequency-domain trained 1D CNN defense model. In particular, the baseline receives normalized fixed length temporal windows, while the defense receives normalized FFT (Fast Fourier Transform) transformed frequency sequences. Also, both models are evaluated with the introduced adversarial attacks. Thus, across five random patient-level splits and 11 perturbation strengths, the result shows that the defense model achieved higher mean attacked macro F1 for all three attack families: 0.676 versus 0.419 for FGSM, 0.483 versus 0.248 for PGD, and 0.482 versus 0.248 for the combined attack. All in all, the main contribution in this essay is a lightweight framework that changes the classifier input from a temporal window to FFT-magnitude features and trains this frequency-domain model with clean, FGSM, and short-PGD samples.*

*The code can be accessed at GitHub (<https://github.com/ArcaneEmerald/ecg-defense-framework>).*

## Keywords:

*Adversarial machine learning, frequency-domain transformation, wearable ECG monitoring, robust neural networks, physiological signal analysis, edge AI, cyber-secure healthcare, signal perturbation defense*

## 1. INTRODUCTION

### 1.2 The Proliferation of Wearable Diagnostics

Wearable devices have advanced significantly, enabling continuous, ambulatory monitoring of critical physiological signals during daily activities rather than restricting patients to controlled

clinical environments. For example, smart wearable devices facilitate ambulatory cardiovascular care by supporting continuous rhythm observation outside clinical settings [1].

More importantly, the scale of wearable devices is no longer theoretical, it has been proved by real numbers. Just in an Apple Heart Study, there have been 419,297 participants enrolled to evaluate a notification algorithm of screening atrial-fibrillation by detecting irregular pulse on wearable devices [2]. Besides, another similar study called the Fitbit Heart Study got 455,699 participants enrolled for atrial-fibrillation detection [3]. However, challenges remain in ambulatory ECG monitoring, primarily because large-scale data collection does not inherently guarantee high-fidelity ground truth annotations [4]. Although there are still many open problems, some studies show that wearable ECG algorithms can be clinically meaningful. For instance, Bumgarner et al. proved that ECG algorithms on wearable devices can be clinically relevant [5]. And Rajpurkar et al. successfully used CNNs for cardiologist-level arrhythmia detection [6]. However, cybersecurity issues are increased as well at the same time. Authentication, confidentiality, integrity and so on have become core security challenges for medical sensing systems in a wireless body area network survey [7]. And another study also emphasis the increasing concerns of data privacy and malicious manipulation risks [8]. Furthermore, the inherent vulnerabilities of wireless transmission expose these systems to adversarial attacks, which can intentionally induce severe model misclassifications.

### **1.3 The Adversarial Threat Vector**

Furthermore, the inherent vulnerabilities of wireless transmission expose these systems to adversarial attacks, which can intentionally induce severe model misclassifications. Malicious attackers can add some small and deliberately generated perturbations to the input raw data so that corresponding trained models will make wrong predictions while the signal still appears plausible to humans. For example, Szegedy et al. find out that deep neural networks are very vulnerable to carefully crafted adversarial perturbations [9]. And Goodfellow et al. show that small worst-case perturbations can lead to misclassification according to their FGSM attack method [10]. Moreover, this kind of attack could result in serious consequences in clinical settings since researchers have proved that adversarial attacks can cause potential safety risks in medical machine learning systems [11]. Particularly in ECG, Han et al. crafted deliberately designed examples of input data for single-lead ECG to fool neural networks [12]. And Ono et al. successfully manipulated diagnosis results by adding adversarial beats into ECG sensor readings [13]. These studies deepen corresponding research into specific ECG security and show that ECG classification systems powered by machine learning algorithms are vulnerable against adversarial attacks, so the need to test ECG classifiers under adversarial conditions and design corresponding defenses stands out.

### **1.4 Ineffective Defense for Wearable Computers**

Several solutions have been proposed for defending adversarial attacks over time, but many of them are not ideal for wearable devices. Traditional security methods, such as encryption and authentication, are inherently incapable of protecting models against carefully crafted perturbations injected directly into the raw physiological input data. The Internet of Medical Things (IoMT) is particularly susceptible to these broader cybersecurity risks [8]. Therefore, some defensive approaches are introduced in adversarial machine learning, such as adversarial training, defensive distillation, detection and so on. Some of them are effective. For instance, Madry et al. showed that adversarial training can significantly improve the robustness of models against adversarial attacks [14]. But some methods are not that effective. For example, Carlini and Wagner demonstrate that defensive distillation does not significantly improve the robustness of neural networks [15]. On top of that, model size is another concern in terms of efficiency of wearable

devices. The parameter count plays a core role in computational efficiency, which is critical given the strict power constraints of wearable devices.

Hence, some models specifically designed for small devices are invented, such as MobileNet that has only 4.2 million parameters and SqueezeNet that achieves accuracy with 50x fewer parameters [16], [17]. These lightweight models reduce model size compared with general-purpose architectures.

### **1.5 Problem Statement & Hypothesis**

Transforming time-series signals into the frequency domain isolates a distinct feature space that is highly effective for analyzing and mitigating adversarial attacks. Therefore, this study presents a lightweight frequency-domain-based method for protecting machine learning models against adversarial perturbations. The Fast Fourier Transform, an efficient way to calculate Fourier Transform on computers, will be used to transform raw physiological signals in temporal domain into frequency domain [18]. This study trains two models, one is a baseline 1D CNN model and the other is a frequency-domain-based adversarial defense 1D CNN model, and they will be attacked by both FGSM and PGD methods, then comparing their separate accuracy in the end. The defense model is expected to have better attacked macro F1 than the baseline under the tested attack settings. This approach is purely software-based and does not require any hardware changes.

## **2. Related Work**

### **2.1 Adversarial Vulnerabilities in 1D Time-Series Data**

For most ECG wearable devices, real-time physiological signals extracted from sensors are usually one-dimensional time-series segments, such as ECG beats. Therefore, one-dimensional convolutional neural networks (1D CNNs) are widely adopted, as they efficiently process physiological signals while maintaining a compact, energy-efficient architecture compared to 2D CNNs [19]. Many prior works have proved that applying deep neural networks to physiological signal detections is effective and efficient. For example, Hannun et al. used deep neural network to successfully classify cardiologist-level arrhythmia in a large ambulatory ECG dataset [20]. Acharya et al. applied a deep CNN to classify heartbeats, showing the increasing popularity of CNN models in heartbeat recognition [21]. Additionally, Kachuee et al. introduced a deep transferable representation of ECG heartbeat classification, which can transfer acquired knowledge to another task [22]. What's more, Ribeiro et al. proved that deep neural networks are able to learn clinically diagnostic patterns from 12-lead ECG recordings, with high accuracy [23]. However, they are not perfect since adversarial attacks can easily corrupt models. Some researchers show that single-lead ECG classifiers are simply affected by adversarial examples while remaining plausible output [12]. Also, Ono et al. demonstrated successful diagnosis manipulation 3-5 times out of 40 attempts during a 2-minute physical ECG experiment [13]. These adversarial attacks not only degrade predictive accuracy but also force models to generate clinically plausible false outputs that easily evade human detection.

### **2.2 Smoothing, Distillation, and Other Defense Techniques**

In order to make the input of a deep learning system more robust, some input defense techniques are used and most of the existing work include approaches of input smoothing and distillation. For instance, Cohen et al. showed that randomized smoothing can provide certified robustness within a defined perturbation radius [24]. However, these methods have their own limitations. For example, Carlini and Wagner showed that defensive distillation can be bypassed when it is evaluated with stronger attacks [15]. Additionally, alternative defensive paradigms, such as input denoising, have been explored to mitigate adversarial perturbations. An ECG baseline wander removal filter based on deep learning called DeepFilter has been introduced by Romero et al. to

address noise in signals [25]. In addition, Sarafan et al. introduced an Ensemble Kalman filter for ECG denoising [26]. These methods address signal noise, but adversarial perturbations are optimized against the classifier, so this study evaluates robustness directly rather than assuming that denoising alone is sufficient. Thus, in this study, a real FFT firstly represents the spectral structure of each ECG window. Then the defense model learns directly from normalized log-magnitude values together with robust training. This experiment checks whether this complete frequency-domain design improves attacked classification while keeping a small CNN.

### 3. Methodology

#### 3.1 Dataset and Preprocessing

The data of this study was created from extracted recordings of ECG devices in the public MIT-BIH Arrhythmia Database (MIT-BIH database) provided by PhysioNet ([www.physionet.org](http://www.physionet.org)), it contains two-channel ambulatory ECG recordings from 47 subjects and these recordings are sampled at 360 samples per second per channel [27]. The preprocessing is as follows: all 48 recordings from record 100 to record 234 were read from the raw folder, using values from .hea and .dat files and annotations from .atr files. The database comprises 47 unique subjects; records 201 and 202 originate from the same patient and are therefore strictly grouped during data partitioning. In addition, the MLII channel was selected when available. Next, each recording was divided into non-overlapping continuous 5-second windows, equal to 1,800 samples. Then a window was labelled abnormal if at least one non-normal annotated beat occurred inside it. Otherwise, it was labelled normal. After that, each window was min-max normalized to  $[-1, 1]$ . What's more, patient groups were separated before window extraction. Thus, in the seed-42 split, 32 training patient groups produced 11,888 train windows and 15 held-out testing patient groups produced 5,415 test windows, with no patient overlap. Keeping patients separate before window creation is important because ECG shape can be specific to one person. If windows from one recording appeared in both datasets, the CNN could partly recognize that recording instead of learning a pattern that transfers to an unseen patient. The 5-second duration is also long enough to contain several beats and short rhythm changes, while it still gives fixed-size inputs for batch training. Non-overlapping windows reduce repeated samples, and only the short remainder at the end of a recording is not used.

#### 3.2 Threat Model Formulation

This study simulates adversarial attack noise by taking advantage of model gradients to perturb normalized ECG windows. For the adversarial threat model, FGSM, PGD, and a combined worst-case FGSM/PGD attack were explicitly implemented to generate targeted perturbations within the ECG pipeline. During our final testing, each model is attacked using its own gradients and evaluated on those same perturbed held-out windows. And both models use the same argmax decision rule for all predictions. Attack strength in this experiment is controlled by epsilon under an L-infinity bound. This experiment evaluates 11 values from 0.005 to 0.50, as shown in Table 1. Here, FGSM uses one gradient step and PGD uses 10 projected steps and retains the highest loss iterate for each window. In addition, the combined method of both FGSM and PGD independently generates FGSM and PGD candidates and retains the higher loss candidate per window, so the combined stress test cannot be weaker than either component merely because of initialization.

**Table 1. Epsilon value used for adversarial attacks.**

| Attack Method | Epsilon( $\epsilon$ ) |   |   |   |   |   |   |   |   |    |    |
|---------------|-----------------------|---|---|---|---|---|---|---|---|----|----|
|               | 1                     | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |

| FGSM &<br>PGD &<br>Combined | 0.005 | 0.01 | 0.02 | 0.03 | 0.05 | 0.08 | 0.10 | 0.15 | 0.20 | 0.30 | 0.50 |
|-----------------------------|-------|------|------|------|------|------|------|------|------|------|------|
|-----------------------------|-------|------|------|------|------|------|------|------|------|------|------|

### 3.3 Feature Extraction in the Frequency Domain

The core mechanism of the defense model is direct learning from normalized FFT-magnitude features, which allows the CNN to learn information from the frequency domain.

The baseline model processes the original normalized 5-second temporal ECG window, establishing a standard benchmark for subsequent comparisons. Secondly, for the defense model, the same temporal window is transformed with a real FFT, converted to magnitude, compressed with  $\log_{1p}$ , and normalized to  $[-1, 1]$ . Then the resulting 901-bin frequency sequence is passed directly into the 1D CNN, which gives the defense model a spectral input representation while the baseline model retains the temporal input representation. The complete 901-bin sequence forms the defense input. Every bin from the real FFT, from 0 Hz to the Nyquist frequency of 180 Hz, remains available to the CNN. The  $\log_{1p}$  operation compresses very large magnitude differences, and the following min-max step keeps the feature range consistent. These steps give a fixed spectral sequence in which the CNN can learn relationships across frequencies.

Figure 1 below shows one ECG window before and after simulated attack noise in both time domain and frequency domain views. It also illustrates how the perturbation changes both the temporal waveform and spectral energy used by the two input representations.

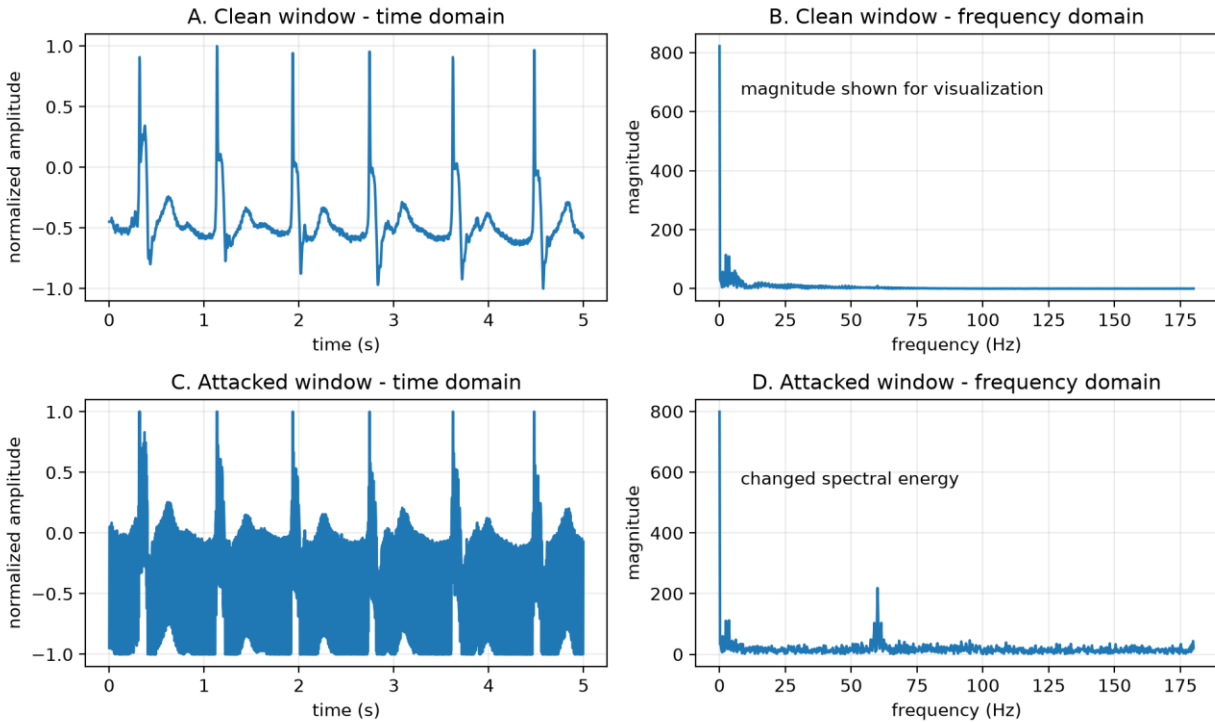


Figure 1. Time and frequency domain comparison of clean and adversarial ECG windows.

### 3.4 Lightweight Deep Learning Architecture

After temporal domain or frequency domain feature preparation, a compact 1D CNN is used for both baseline model and defense model. Compact 1D CNNs are highly suitable for these ECG signals, as they efficiently process one-dimensional physiological data with a significantly smaller parameter footprint than vision-based 2D CNNs [19]. This CNN uses three convolutional blocks. And each block contains a 1D convolutional layer, batch normalization, and ReLU activation. The

convolution uses kernel size 7 with padding 3, and max pooling is applied after the first two blocks, then adaptive average pooling compresses the final feature map into a 64-dimensional representation. Both models contain 18,498 trainable parameters and use the same binary linear head. Their only difference of input domain is that the baseline model's encoder receives 1,800 temporal domain samples while the defense model's encoder receives 901 frequency domain FFT-magnitude bins. The Figure 2 below shows the complete architecture of 1D CNN used for both baseline model and defense model.

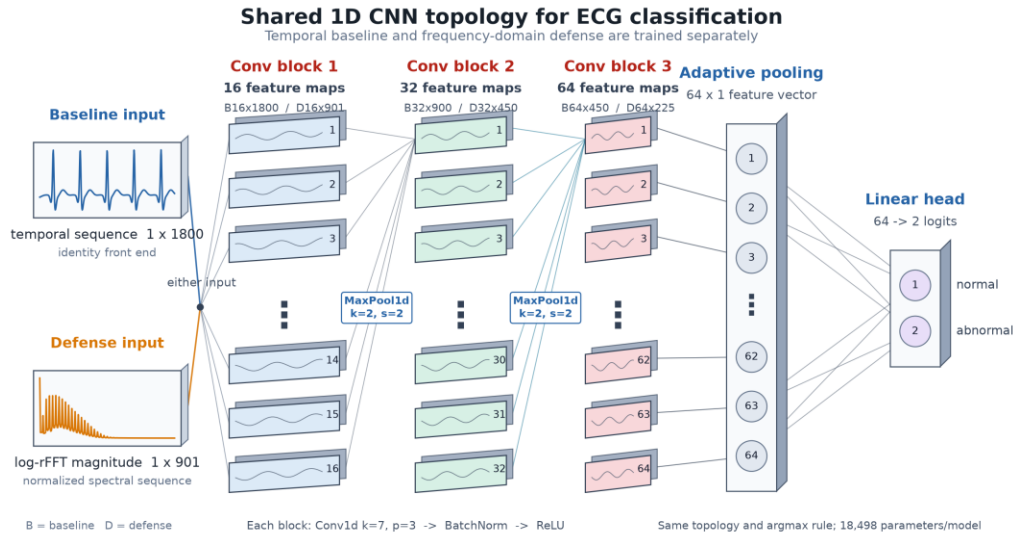


Figure 2. 1D CNN architecture for both temporal domain baseline model and frequency domain defense model.

### 3.5 Summary and Experimental Workflow

First of all, our workflow begins with the MIT-BIH dataset, which is downloaded and read through WFDB. Then patient groups are divided into 70/30 train/test datasets. Within the training patients, approximately 20% are held out as an inner validation set for hyperparameter and epoch selection. After these patient-level splits, continuous 5-second ECG windows are extracted and normalized. This order prevents windows from the same patient appearing across training, validation, and test sets. The baseline model is trained exclusively on clean temporal-domain windows. Conversely, the defense model is trained on FFT-magnitude features using a composite loss function (50% clean, 25% FGSM, and 25% three-step PGD), with both training attacks bounded by an epsilon of 0.10. And baseline checkpoints are selected by clean validation macro F1 as well as defense checkpoints are selected by the mean of clean, FGSM, and PGD validation macro F1 at epsilon 0.05. This validation rule selects checkpoints that balance clean and attacked macro F1.

In this study, adversarial training is included in the defense because FFT changes the input representation but does not guarantee resistance to gradient-based perturbations by itself. FGSM and short PGD samples allow the optimizer to see the local worst-case directions that are included in testing. The baseline remains clean trained as the unprotected reference. Therefore, the comparison evaluates the complete defense framework, which combines frequency-domain features and robust training. Thirdly, both models are evaluated on held-out patient records under clean input and model-targeted FGSM, PGD, and combined FGSM/PGD attacks. In this step, learning rates of  $1e-4$ ,  $3e-4$ ,  $1e-3$ , and  $3e-3$  are compared on the inner validation patients. And the selected shared rate in this essay is  $1e-3$ . Next, the complete evaluation is repeated for five patient-split seeds, 42 through 46, and attack plots report the mean with one-standard-deviation bands.

Finally, we analyzed both robustness and efficiency for both models based on our collected and calculated metrics.

#### 4. Experimental Setup

##### 4.1 Hardware and Software Environment

All of the experiments were conducted on a desktop-based environment rather than server-based environment since the computational demand of this study is not high, and only pure software was used as well. A description of the hardware as well as software configurations is provided next.

Model training and evaluation were executed on an Apple M3 Pro processor equipped with 18 GB of unified memory. For software, the running Python version is 3.12.12 and a virtual environment of Python is also created in code folder; the installed packages are torch, numpy, pandas, scikit-learn, matplotlib, wfdb. These packages were used for different parts: the torch package was used to build and train the neural network models, calculate gradients, and generate adversarial examples; the numpy package was used for numerical calculation and ECG signal preprocessing; the pandas package was used to store and organize experimental results, such as training history, robustness metrics, and summary tables; the scikit-learn package was used to calculate evaluation metrics, including accuracy, precision, recall, and macro F1 score; the matplotlib package was used to generate waveform, CNN architecture, attack-strength, defense-comparison, and learning-rate figures; the wfdb package was used to download and read ECG records and annotations from the MIT-BIH Arrhythmia database. Besides, this study also used Git to manage the codebase and version control. When it comes to IDEs, this study used PyCharm and VS code.

What's more, all experiment settings are stored in src/config.py, including the patient-split seeds, learning-rate options, attack strengths, and training settings. The command “python run.py all” prepares the selected splits, tunes the learning rate, trains the five repeated experiments, evaluates each attack, runs the benchmark, and creates Figures 1 and 3-5.

##### 4.2 Evaluation Metrics

In order to have a full evaluation, this study measures both classification robustness and computational efficiency for ECG window classification.

The main metrics are accuracy, precision, recall, macro F1, attacked macro F1, estimated FLOPs, latency, and parameter count. In this study, accuracy indicates overall correctness of predictions for both baseline and defense models, the formula of accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision indicates whether predicted abnormal windows are truly abnormal since we defined the label of abnormal signal to 1 in our models, the formula of precision is:

$$Precision = \frac{TP}{TP + FP}$$

Recall indicates whether truly abnormal windows were all found, the formula of recall is:

$$Recall = \frac{TP}{TP + FN}$$

F1 balances precision and recall, the formula of F1 is:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Macro F1 averages class-level F1 equally, which is why it becomes the main classification metric in this study for this binary ECG classification task, the formula of binary Macro F1 is:

$$Macro\ F1 = \frac{F1_0 + F1_1}{2}$$

Mentioned above TP means true positive, TN true negative, FP false positive, and FN false negative.

In addition, estimated FLOPs, latency, and parameter count are measured because robustness is not the only concern for a wearable classifier. In this study, FLOPs estimates the arithmetic operations required for one forward pass; it is not a direct measurement of energy. Latency is the average number of milliseconds required to classify one window under the stated desktop benchmark, while throughput reports windows per second. Parameter count describes the trainable model size. Actual battery cost still needs measurement on wearable hardware.

## 5. Results and Analysis

### 5.1 Hyperparameter Tuning and Cross-Validation

In this study, learning-rate tuning compared  $1e-4$ ,  $3e-4$ ,  $1e-3$ , and  $3e-3$  using only seed-42 inner validation patients. Therefore, Figure 3 below shows validation accuracy, not test accuracy. This figure shows that the baseline model's highest validation accuracy was 0.953 at 0.003, while the defense model reached 0.918 at 0.001. In addition, a shared rate of  $1e-3$  was selected because it is the defense model optimum and keeps baseline model's validation accuracy at 0.948, close to its maximum, without using test patients for selection. What's more, to reduce dependence on one train/test partition, the final adversarial evaluation was repeated over five random patient-level splits with seeds 42, 43, 44, 45, and 46. Every outer split contained 32 training patient groups and 15 test patient groups, with a disjoint inner validation subset. Also, Figures 4 and 5 therefore report mean macro F1 and standard deviation across five independently reselected patient splits.

The inner validation patients are used only for learning-rate and checkpoint selection. They are not included in the final test calculation. For every new seed, patient groups are selected again before any window is extracted, so the mean and standard deviation represent five different held-out patient sets instead of five repeats on one fixed test set.

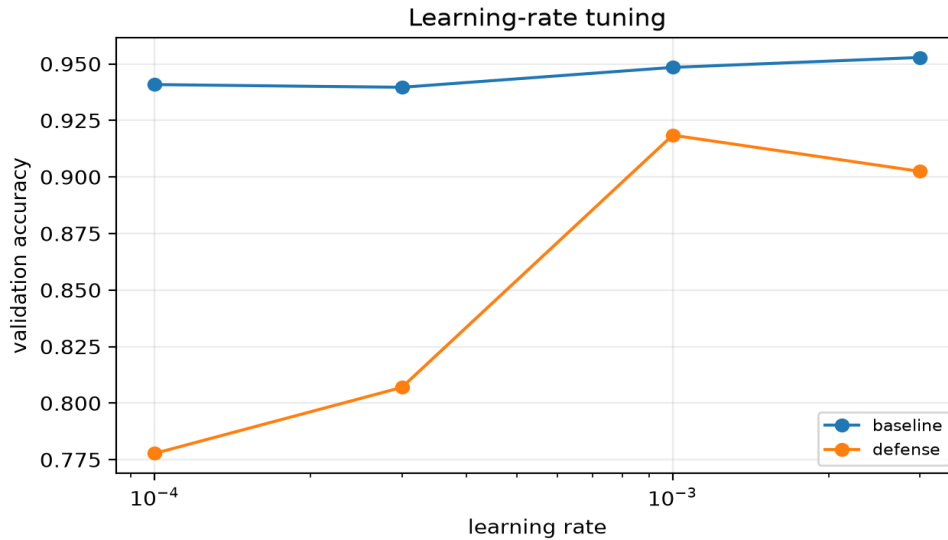


Figure 3. Learning-rate tuning for temporal baseline and FFT-feature model.

### 5.2 Assessment of the Vulnerability of the Unprotected Model

The unprotected temporal baseline was evaluated under all three attack families across five patient-level splits. And its clean macro F1 was  $0.765 \pm 0.048$ . Averaged across the 11 epsilon values, attacked macro F1 fell to  $0.419 \pm 0.085$  under FGSM,  $0.248 \pm 0.064$  under PGD, and  $0.248 \pm 0.064$  under the combined stress test. In addition, Figure 4 shows the corresponding mean curves and

standard-deviation bands. In terms of FGSM, baseline model's macro F1 was  $0.522 \pm 0.113$  at epsilon 0.005,  $0.405 \pm 0.127$  at epsilon 0.10, and  $0.355 \pm 0.021$  at epsilon 0.50. Furthermore, the apparent plateauing of the metric at large epsilon values requires careful interpretation via confusion matrices, as overwhelming perturbations frequently force the baseline model to collapse into single-class prediction rather than recover its accuracy. We can also find that the iterative attacks caused the larger reduction. For example, at epsilon 0.05, baseline model's macro F1 was  $0.215 \pm 0.087$  for PGD and  $0.215 \pm 0.087$  for the combined setting. In addition, the combined attack is a worst-case union: FGSM and PGD candidates are generated independently, and the higher-loss candidate is used for each window. The standard-deviation bands are also important here. A wide band means that the result changes with the patients assigned to a split, especially when one split has a different balance of abnormal windows. At large epsilon values, a flat macro F1 curve does not mean that the model has recovered. The confusion counts are checked because predicting mainly one class can also create a flat curve.

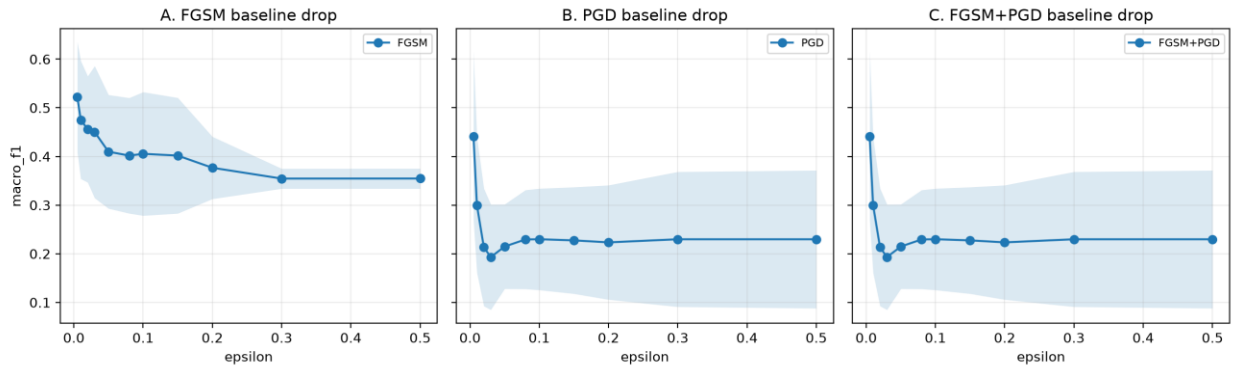


Figure 4. Unprotected baseline vulnerability under FGSM, PGD and combined FGSM/PGD attacks.

### 6.3 Defense Efficacy

When it comes to defense efficacy, Figure 5 below compares adversarial attacks generated against and evaluated on each model. It shows that the defense model's clean macro F1 was  $0.726 \pm 0.040$ , compared with  $0.765 \pm 0.048$  for the baseline. In addition, this clean difference is reported separately because the adversarial robustness should not be inferred from clean accuracy alone. These adversarially attacked results show the benefit of the complete FFT-magnitude plus robust-training framework. What's more, this figure also indicates that averaged across all 11 FGSM strengths, defense model's macro F1 was  $0.676 \pm 0.032$ , compared with  $0.419 \pm 0.085$  for the temporal baseline model. For instance, at epsilon 0.10, the defense model achieved  $0.696 \pm 0.034$ , while the baseline model achieved  $0.405 \pm 0.127$ . Notably, the defense model mean exceeded the baseline model at 11 of 11 plotted FGSM strengths. Additionally, the iterative PGD results show the larger advantage in this figure. For example, averaged across the epsilon grid, defense model's macro F1 was  $0.483 \pm 0.043$  under PGD versus  $0.248 \pm 0.064$  for the baseline model, and  $0.482 \pm 0.044$  under the combined attack versus  $0.248 \pm 0.064$ . And at epsilon 0.05, the defense model reached  $0.550 \pm 0.038$  under PGD and  $0.549 \pm 0.038$  under the combined attack.

Overall, the defense model achieved higher mean adversarial attacked macro F1 for all three attack families and exceeded the baseline model's mean curve at 10 of 11 PGD strengths and 10 of 11 combined-attack strengths. These results support the complete frequency-domain defense framework under the tested digital threat model. But they do not prove that FFT magnitude alone causes the improvement, because robust training is also part of the defense.

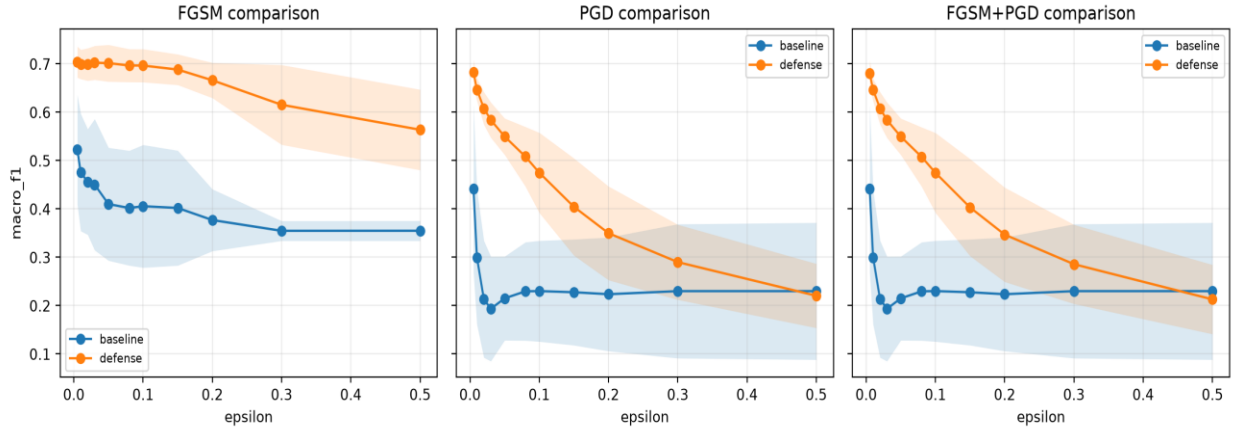


Figure 5. Baseline model versus frequency defense model under adversarial attacks.

### 5.4 Computational Overhead

When it comes to computational overhead, Table 2 below indicates the computational comparison. It shows that both models have 18,498 trainable parameters since they use the same CNN structure. Also, the temporal domain baseline model requires approximately 19,757,056 operations per window while the frequency domain defense model requires approximately 9,976,204, including an estimated 97,324 operations for the FFT. On top of that, its convolutional sequence is shorter because the real FFT produces 901 magnitude bins from 1,800 temporal samples.

In addition, measured latency was 0.3793 ms per window for the baseline model and 0.4750 ms for the defense model. And corresponding throughput was 2636.63 and 2105.07 windows per second. What's more, for example, these synchronized measurements include FFT feature extraction in the defense forward pass. Therefore, the defense model keeps the same parameter count and a small checkpoint while adding only a little bit frequency-transform cost. So, the measurements support a lightweight software implementation, although real wearable deployment would still require device-specific energy and latency testing.

Table 2. Computational overhead comparison between baseline and defense models.

| Model    | Parameters | Estimated FLOPs | Checkpoint Size (MB) | Latency (ms/window) | Windows (/s) |
|----------|------------|-----------------|----------------------|---------------------|--------------|
| Baseline | 18,498     | 19,757,056      | 0.0799               | 0.3793              | 2636.63      |
| Defense  | 18,498     | 9,976,204       | 0.0799               | 0.4750              | 2105.07      |

## 6. Discussion

### 6.1 Spectral Feature Mechanism

We can tell from our experiments that FFT-magnitude features and robust training have separate jobs in the defense. First of all, the spectral representation summarizes how energy is distributed across the 5-second window, so the CNN depends less on the exact position of one beat. In addition, log scaling keeps a few large spectral peaks from dominating all other bins. Robust training then exposes this frequency-domain model to both one-step FGSM and iterative PGD directions before the final test. This second step is necessary because changing the representation alone does not remove gradients from the classifier. Figure 5 therefore reports the performance of the complete defense framework. The higher attacked macro F1 supports this combined design under the tested white-box setting, but it does not prove that frequency-domain features alone are

responsible for the whole gain. A matched temporal adversarial-trained model and a clean-trained frequency model are still needed to separate both contributions.

## 6.2 Deployment of the Project to Real-World Settings

In practice, this lightweight CNN architecture can be seamlessly deployed as a software module within an existing ECG classification pipeline. The baseline model will receive normalized temporal windows directly while the defense model will calculate log-scaled FFT magnitude and classify the resulting frequency sequence. Also, this transform requires no hardware change and uses the same argmax output rule as the baseline. However, deployment of this project should still be tested under device energy constraints and real-world sensor corruption.

## 6.3 Current Vulnerabilities of the System

Although the complete frequency-domain based defense model improves the tested adversarial attack averages, this current system still has several limits and corresponding directions for improvement. First of all, the adversarial attacks are digitally simulated. They provide controlled, model-targeted evaluation, but they do not fully represent wearable conditions where corruption may originate in electrodes, analogue front ends, wireless transmission, or malicious software. Thus, future work should consider, and test physically injected perturbations, corrupted sensor streams, and ECG recorded from multiple wearable devices.

**Secondly**, the reported comparison evaluates an unprotected temporal domain based baseline model against a frequency domain based defense model that changes both input domain and training objective. This is appropriate for testing the complete framework, but it does not isolate the causal contribution of FFT magnitude from adversarial training. Therefore, future work should add a temporal adversarial trained control and a frequency-only training step, and should evaluate adaptive attackers that know the complete feature extractor and training procedure.

**Thirdly**, performance varies across patient splits, and a 5-second window is labelled abnormal when any abnormal beat appears inside it. Thus, future work should test larger ECG datasets, patient-stratified label distributions, multi-class rhythm labels, and longer contextual models while continuing to keep every patient's window in only one data group.

## 7. Conclusion

### 7.1 Summary of Contributions

This study proposed and evaluated a lightweight frequency-domain defense framework for ECG classification under FGSM, PGD, and combined attacks. First of all, the temporal domain based baseline model is trained on normalized 5-second multi-beat ECG windows while the frequency domain based defense model is trained directly on normalized log-scaled FFT-magnitude features with clean, FGSM, and short-PGD losses. Both models use the same 18,498-parameter CNN and argmax decision rule, and patients remain separated across data partitions. Also, learning-rate tuning selected  $1e-3$ , and five random patient splits produced mean and standard-deviation curves through epsilon 0.50. As a result, the frequency domain based defense model achieved higher mean attacked macro F1 for all three attack families.

### 7.2 Future Work

Future work should test physically injected and adaptive attacks in real clinical wearable settings. It should also add matched temporal adversarial-training and clean frequency-domain controls, so the separate effect of representation and training can be measured. Larger patient groups, multi-class rhythm labels, and signals collected from different wearable devices should be considered as well. In addition, repeated external validation is needed to check whether the same robustness appears beyond MIT-BIH. Finally, latency and energy should be measured on an actual wearable

processor rather than inferred only from a desktop benchmark. These steps can answer the remaining questions about patient generalization, causal mechanism, and deployment.

## Reference

K. Bayoumy *et al.*, “Smart wearable devices in cardiovascular care: where we are and how to move forward,” *Nat. Rev. Cardiol.*, vol. 18, no. 8, pp. 581–599, Aug. 2021, doi: 10.1038/s41569-021-00522-7.

M. V. Perez *et al.*, “Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation,” *N. Engl. J. Med.*, vol. 381, no. 20, pp. 1909–1917, Nov. 2019, doi: 10.1056/NEJMoa1901183.

S. A. Lubitz *et al.*, “Detection of Atrial Fibrillation in a Large Population Using Wearable Devices: The Fitbit Heart Study,” *Circulation*, vol. 146, no. 19, pp. 1415–1424, Nov. 2022, doi: 10.1161/CIRCULATIONAHA.122.060291.

D. Liaqat, R. Wu, S. Liaqat, E. de Lara, A. Gershon, and F. Rudzicz, “The Ground Truth Trade-Off in Wearable Sensing Studies,” Jan. 07, 2020, *arXiv*: arXiv:2001.09738. doi: 10.48550/arXiv.2001.09738.

J. M. Bumgarner *et al.*, “Smartwatch Algorithm for Automated Detection of Atrial Fibrillation,” *J. Am. Coll. Cardiol.*, vol. 71, no. 21, pp. 2381–2388, May 2018, doi: 10.1016/j.jacc.2018.03.003.

P. Rajpurkar, A. Y. Hannun, M. Haghighpanahi, C. Bourn, and A. Y. Ng, “Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks,” Jul. 06, 2017, *arXiv*: arXiv:1707.01836. doi: 10.48550/arXiv.1707.01836.

S. Movassaghi, M. Abolhasan, J. Lipman, D. Smith, and A. Jamalipour, “Wireless Body Area Networks: A Survey,” *IEEE Commun. Surv. Tutor.*, vol. 16, no. 3, pp. 1658–1686, 2014, doi: 10.1109/SURV.2013.121313.00064.

F. Alsubaei, A. Abuhussein, and S. Shiva, “Security and Privacy in the Internet of Medical Things: Taxonomy and Risk Assessment,” in *2017 IEEE 42nd Conference on Local Computer Networks Workshops (LCN Workshops)*, Oct. 2017, pp. 112–120. doi: 10.1109/LCN.Workshops.2017.72.

C. Szegedy *et al.*, “Intriguing properties of neural networks,” Feb. 19, 2014, *arXiv*: arXiv:1312.6199. doi: 10.48550/arXiv.1312.6199.

I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” Mar. 20, 2015, *arXiv*: arXiv:1412.6572. doi: 10.48550/arXiv.1412.6572.

S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, “Adversarial attacks on medical machine learning,” *Science*, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019, doi: 10.1126/science.aaw4399.

X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath, “Adversarial Examples for Electrocardiograms,” Jun. 04, 2019, *arXiv*: arXiv:1905.05163. doi: 10.48550/arXiv.1905.05163.

T. Ono, T. Sugawara, J. Sakuma, and T. Mori, “Application of Adversarial Examples to Physical ECG Signals,” Aug. 20, 2021, *arXiv*: arXiv:2108.08972. doi: 10.48550/arXiv.2108.08972.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards Deep Learning Models Resistant to Adversarial Attacks,” presented at the International Conference on Learning Representations, Feb. 2018. Accessed: Jun. 26, 2026. [Online]. Available: <https://openreview.net/forum?id=rJzIBfZAb>

N. Carlini and D. Wagner, “Towards Evaluating the Robustness of Neural Networks,” in *2017 IEEE Symposium on Security and Privacy (SP)*, May 2017, pp. 39–57. doi: 10.1109/SP.2017.49.

A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 17, 2017, *arXiv*: arXiv:1704.04861. doi: 10.48550/arXiv.1704.04861.

F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size,” Nov. 04, 2016, *arXiv*: arXiv:1602.07360. doi: 10.48550/arXiv.1602.07360.

J. W. Cooley and J. W. Tukey, “An algorithm for the machine calculation of complex Fourier series,” *Math. Comput.*, vol. 19, no. 90, pp. 297–301, 1965, doi: 10.1090/S0025-5718-1965-0178586-1.

S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, “1D convolutional neural networks and applications: A survey,” *Mech. Syst. Signal Process.*, vol. 151, p. 107398, Apr. 2021, doi: 10.1016/j.ymssp.2020.107398.

A. Y. Hannun *et al.*, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nat. Med.*, vol. 25, no. 1, pp. 65–69, Jan. 2019, doi: 10.1038/s41591-018-0268-3.

U. R. Acharya *et al.*, “A deep convolutional neural network model to classify heartbeats,” *Comput. Biol. Med.*, vol. 89, pp. 389–396, Oct. 2017, doi: 10.1016/j.combiomed.2017.08.022.

M. Kachuee, S. Fazeli, and M. Sarrafzadeh, “ECG Heartbeat Classification: A Deep Transferable Representation,” in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, Jun. 2018, pp. 443–444. doi: 10.1109/ICHI.2018.00092.

A. H. Ribeiro *et al.*, “Automatic diagnosis of the 12-lead ECG using a deep neural network,” *Nat. Commun.*, vol. 11, no. 1, p. 1760, Apr. 2020, doi: 10.1038/s41467-020-15432-4.

J. Cohen, E. Rosenfeld, and Z. Kolter, “Certified Adversarial Robustness via Randomized Smoothing,” in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, May 2019, pp. 1310–1320. Accessed: Jun. 28, 2026. [Online]. Available: <https://proceedings.mlr.press/v97/cohen19c.html>

F. P. Romero, D. C. Piñol, and C. R. V. Seisdedos, “DeepFilter: an ECG baseline wander removal filter using deep learning techniques,” Jan. 12, 2021, *arXiv*: arXiv:2101.03423. doi: 10.48550/arXiv.2101.03423.

S. Sarafan *et al.*, “A Novel ECG Denoising Scheme Using the Ensemble Kalman Filter,” Jul. 24, 2022, *arXiv*: arXiv:2207.11819. doi: 10.48550/arXiv.2207.11819.

G. B. Moody and R. G. Mark, “The impact of the MIT-BIH Arrhythmia Database,” *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May 2001, doi: 10.1109/51.932724.