Faculty of Medicine

SYSTEMS BIOLOGY

VU Faculty of Medicine

Dalia Vektarytė

Master's thesis

ŠIRDIES PŪPSNIŲ POŽYMIŲ IŠSKYRIMO IŠ TRIUKŠMINGŲ EKG SIGNALŲ GERINIMAS NAUDOJANT GILIOJO MOKYMOSI ALGORITMU

ENHANCING HEARTBEAT FEATURE EXTRACTION FROM NOISY ECG SIGNALS USING DEEP LEARNING ALGORITHMS

Supervisor:	doc dr. Jolita Bernatavičienė
Head of department:	prof. (HP) dr. Algirdas Utkus

Vilnius, 2026

Student's email dalia.vektaryte@stud.mf.vu.lt

CONTENTS

LIST OF ABBREVIATIONS

AFDB – MIT-BIH Atrial Fibrillation Database

AFib – atrial fibrillation

AFL – atrial flutter

AT – atrial tachycardia

BW – baseline wander

CinC2017 – PhysioNet/Computing in Cardiology Challenge 2017 database

CNN – convolutional neural network

DDPMs – diffusion probabilistic models

DL – deep learning

DNNs – deep neural networks

DWT – discrete wavelet transform

ECG – electrocardiogram

EM – electrode motion noise

FE – feature enhancement

FN – false negative

FP – false positive

GAP – global average pooling

GRU – gated recurrent unit

HNF – Half-Normalized Filter

LSTM – long short-term memory

MA – muscle artifact

MAD – maximum absolute difference

MSConv – multiscale convolution

N – normal rhythm or other non-AF rhythm

NSTDB – MIT-BIH Noise Stress Test Database

PAT/NOD – paroxysmal atrial tachycardia / nodal rhythm

PRD – percentage root-mean-square difference

ReLU – rectified linear unit

SE – squeeze-and-excitation

SHDB-AF – Saitama Heart Database Atrial Fibrillation

SNR – signal-to-noise ratio

SSD – sum of squared differences

STFT – short-time Fourier transform

TN – true negative

TP – true positive

INTRODUCTION

Cardiovascular diseases are the leading cause of death worldwide. It was estimated that they were responsible for about a third of all global deaths (Lindström et al. 2022). Within this burden, Atrial fibrillation (AFib) is one of the most commonly diagnosed arrhythmias. It can be dangerous because it can be asymptomatic (Zimetbaum, 2017). It also appears in transient paroxysmal episodes that further complicate the diagnosis (Staerk et al., 2017).

Undiagnosed and untreated AFib can be dangerous, because it can lead to fatal outcomes. One of the possible complications is an ischemic stroke, that happens because ineffective atrial contraction promotes blood stasis in the atria and facilitates thrombus formation. In a Lithuanian study, Masiliūnas et al. (2022) found that 40.8% of ischemic stroke patients had coexisting AFib. This highlights that early detection and timely treatment of AFib is important. One possible solution for earlier diagnosis is the use of Holter monitors and deep learning (DL) algorithms that enable recording of electrocardiograms (ECG) and can improve likelihood of timely detection.

In recent years, many DL models have been proposed for automatic AFib detection with promising results (Liang et al., 2024). However, their practical applicability is limited due to several issues that remain unaddressed. Published studies often use different methodological designs, evaluation metrics and ECG window lengths, which makes fair comparison between results difficult. In addition, wearable ECG recordings are frequently corrupted by motion artifacts and poor electrode contact that can degrade classification accuracy which is often not addressed in detection research.

Signal denoising is an important part of robust wearable AFib detection. Detection research use relatively clean datasets and traditional denoising methods that are insufficient for complex noise present in Holter recordings. Recent denoising papers have reported that diffusion-based methods can be affecting ECG signal reconstruction (Li et al., 2024). However, their value for AFib classification, has not been examined. This is a knowledge gap as denoising approaches are advancing, but their impact on AFib detection has not yet been evaluated.

Therefore, the aim of this master thesis is to investigate the feasibility of AFib detection models on single-lead ECG signals using deep neural networks (DNNs) and to evaluate the impact of signal window length, presence of noise and DL denoising model impact on detection results.

Accordingly, this thesis aims to answer the following questions:

1. To identify the advantages and disadvantages of selected deep learning models for single-lead AFib detection according to their performance across different datasets.
2. To evaluate the impact of different window length on AFib classification results.
3. To evaluate the impact of different loss functions on AFib classification results.
4. To evaluate the influence of noise on AFib detection accuracy and the influence of diffusion-based denoising method on AFib detection performance.

Parts of the research presented in this thesis were disseminated at two national scientific conferences. The work was presented at the 14th Young Scientists' Conference "Fizinių ir technologijos mokslų tarpdalykiniai tyrimai", held at the Lithuanian Academy of Sciences in Vilnius on 17 April 2026. Also, a conference paper related to this thesis was published and presented at the 7th Conference "Lietuvos magistrantų informatikos ir IT tyrimai", held at the Lithuanian Academy of Sciences in Vilnius on 6 May 2026 (Vektarytė et al., 2026).

LAY DESCRIPTION

AFib is one of the most common heart rhythm disorders. It can be dangerous because it often causes no obvious symptoms and can go undetected. When AFib is undetected not treated in time, it can increase the risk of stroke and mortality. Holter monitors can record heart activity for long periods of time and when used together with DL algorithms designed for automated episode detection, can help diagnose AFib earlier. However, it is difficult to compare the performance of published DL models as they use different methodologies and report different metrics. What is more, when patients wear ECG devices, shifts in electrodes and body movement can introduce noise in the recordings that are not addressed in detection research and in real world applications could reduce AFib detection accuracy.

Therefore, this thesis focuses on improving AFib detection from wearable ECG recordings so that it is more reliable in real-world use. This work investigates how different published deep learning models perform on AFib detection, using the same methodology and evaluation metrics for all of them. It also explores if choice of used ECG window length and loss function improves accuracy of detection. And lastly, it evaluates the impact of noise and if new denoising methods can remove noise effectively and improve AFib detection. By addressing these questions, the thesis aims to pave the way for the development of more robust AFib detection pipelines for wearable monitoring.

LITERATURE REVIEW

1.1. THE STRUCTURE AND FEATURES OF THE ELECTROCARDIOGRAM

Heart rhythm is defined as coordinated electrical activity that causes contractions of the chambers of the heart. Electrical impulses are generated in the atrium by specialized cells in the sinoatrial node (Fig. 1). These impulses travel further through the atria and cause atrial contraction. Then they travel to the atrioventricular node, conduction slows down as the blood fills the ventricles before they contract and the blood is pushed into the lungs and the rest of the body (Wagner, 2008). Proper conduction of electrical impulses allows the heart to pump blood efficiently.

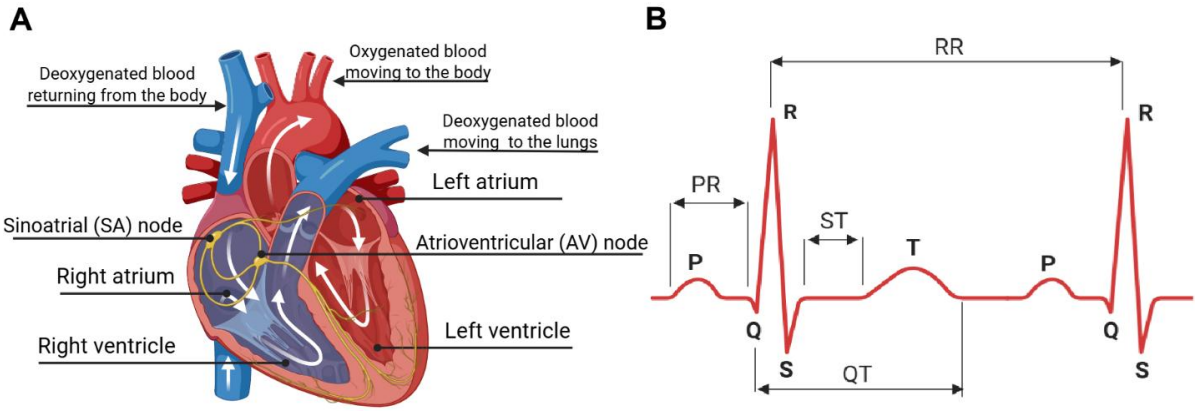

Figure 1. Cardiac structure and ECG signal representation. A - shows anatomical structure of the heart and the direction of blood flow. B - illiustrates standard deflections (P, QRS, T) and key time intervals (PR, ST, QT, RR) of ECG waveform. Figure was created with BioRender.com.

Different components of the ECG recording correspond to specific phases of the heartbeat. The ECG is divided into P, Q, R, S, T and sometimes an additional U wave (Wagner, 2008). The P wave reflects atrial depolarization that results in contraction of the atria. The QRS complex shows ventricular depolarization, when the ventricles are activated to contract and push the blood out into the circulatory system. It is followed by the T wave, which represents ventricular repolarization as the heart returns to its resting membrane potential before the next beat (Wagner, 2008). The time intervals between these waves also provide important information about cardiac conduction. For example, the PR interval represents the time needed for the impulse to travel from the atria to the ventricles, the ST segment corresponds to the period when the ventricles are depolarized, and the QT interval represents the total time of ventricular activity. Also, the R–R interval measures the

time between two consecutive heart beats and is used to evaluate heartbeat regularity (Wagner, 2008). Thus, different components of the ECG and time intervals between them provide important information about electrical signal conduction in the heart and fluctuations in heart rhythm.

ECG is recorded when electrodes are placed on the skin and they detect small voltage changes generated by the heart. The comparison of electrical potentials between two electrodes, displayed graphically over time, is called a lead. When the electrical impulse moves toward the positive electrode of a lead, the ECG shows an upward deflection, and movement away from it produces a downward deflection (Wagner, 2008). Hospitals typically use 12-lead recording measured by 10 electrodes placed on the chest and limbs. However, if continuous monitoring is needed, patients use Holter monitors that can only detect one or two leads (Sibomana et al., 2025). This provides a more limited amount of information for detection of diseases and is more susceptible to noise, because you cannot rely on multiple leads for detection if one of them is corrupted due to electrode shift.

1. 2. CHARACTERISTICS OF ATRIAL FIBRILLATION

AFib is a cardiac arrhythmia in which electrical signal propagation from the atria is disrupted. It can be caused by fibrosis or other structural changes in the atrial myocardium that alter or block normal electrical conduction (Ioannidis et al., 2021). Because of this disturbance of signal propagation, we see small disorganized electrical impulses called fibrillatory waves instead of the P wave in the ECG (Ioannidis et al., 2021). As signal travels further, the timing of the heartbeats becomes highly irregular, which is visible in the R–R intervals in the ECG (Fig. 2).

During clinical diagnosis of AFib, characteristic features need to be present in the ECG recording for a sufficient duration of time. According to previous European Society of Cardiology guidelines, AFib is diagnosed if the arrhythmia is present for at least 30 seconds (Van Gelder et al., 2024). However, this criterion introduces a diagnostic gap for patients with shorter, paroxysmal AFib episodes that do not meet the minimum duration threshold. Currently, hospitals currently use standard 10 second ECG records for diagnosis. And because AFib episodes can occur and resolve spontaneously (Chiang et al., 2012), ambulatory Holter monitors are used to monitor cardiac

activity over a 24-hour period if the patient is suspected to have an underlying rhythm disorder (Zimetbaum, 2017). Therefore, the signal window length used for AFib detection remains an important methodological and clinical question, especially if AFib appears in short recurring episodes that can still lead to adverse outcomes if undetected.

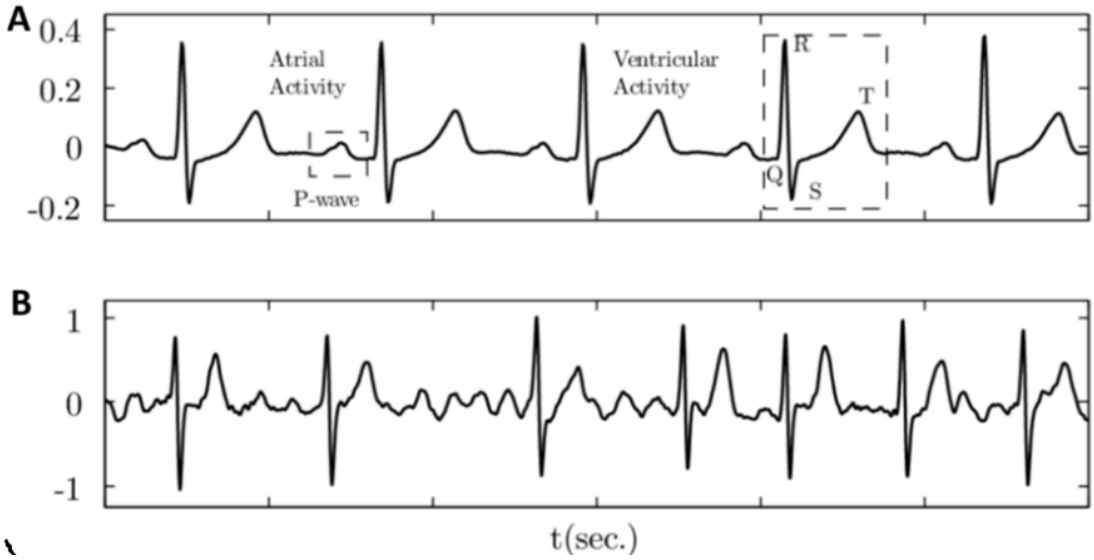

Figure 2. Representative ECG signals illustrating normal sinus rhythm and atrial arrhythmias. A - shows normal sinus rhythm with clearly identifiable P waves. B - atrial fibrillation (AFib), characterized by the absence of distinct P waves, and highly irregular R–R intervals (Llinares and Igual, 2011).

1. 3. NOISE IN THE ELECTROCARDIOGRAPHIC RECORDINGS

ECG recordings are often polluted with noise that degrades signal quality and makes signal analysis more difficult. Some types of noise are quite homogeneous, and have a predictable nature, and are therefore easier to remove (Fig. 3). For example, baseline wander (BW) shows up as shifts of the isoelectric line and is mainly caused by breathing (Tripathi et al., 2022) and power-line interference is caused when an ECG recording device detects electrical signals from the surrounding environment and shows up at a frequency of 50 Hz in Lithuania.

However, wearable devices are also heavily affected by more challenging types of noise that overlap with the frequency spectrum of the ECG and obscure important features. Muscle artifacts (MA) are introduced due to muscles contracting during movement and electrode motion (EM)

noise appears in the recording due to shifts of the electrode on the skin (Tripathi et al., 2022). In a clinical setting, movement related noise can be minimized by instructing the patient to remain still, but wearable devices are worn during daily activities, and this type of noise cannot be fully avoided. In addition, this movement related noise is especially difficult to remove from the ECG signal due to its unpredictability and the fact that it overlaps with the clinically important features.

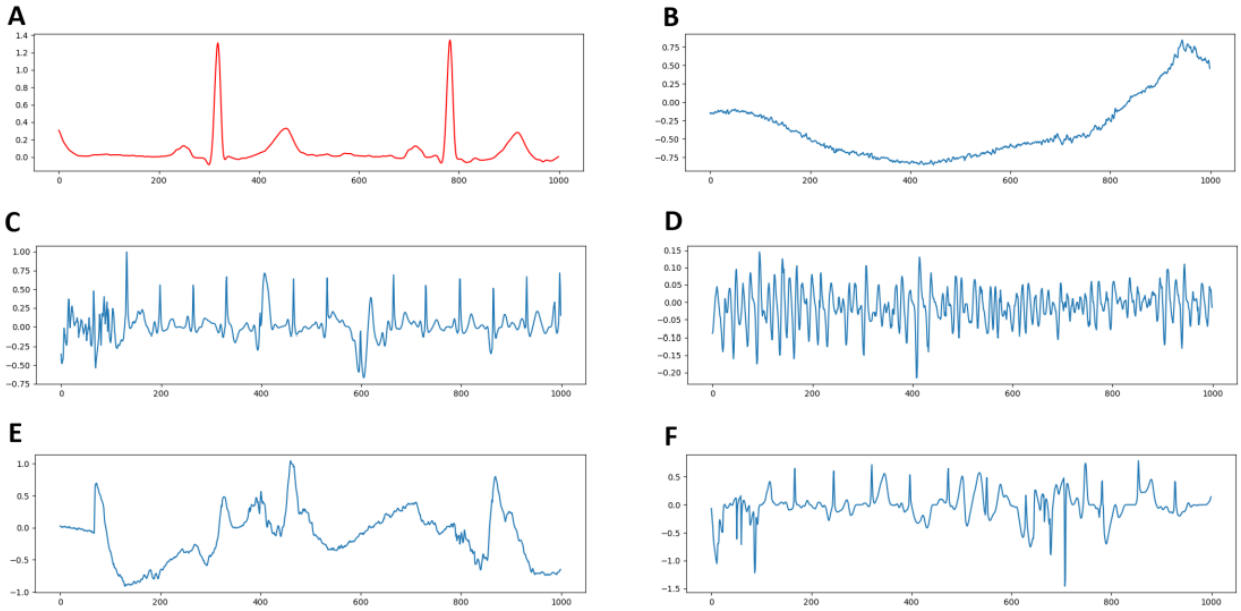

Figure 3. Representative ECG signals illustrating common noise sources in wearable ECG recordings. A – clean ECG signal with clearly identifiable cardiac morphology. B – baseline wander. C – powerline interference. D – electrode motion noise. E – motion artifacts. F – combined effect of various noises (Hu et al., 2024).

1.4. DATASETS RELEVANT TO ATRIAL FIBRILLATION DETECTION

AFib classifiers are commonly trained and evaluated using publicly available annotated ECG databases (Table 1). Since the PhysioNet/Computing in Cardiology Challenge 2017 (Clifford et al., 2017), the CinC2017 database has been frequently used benchmark dataset. It contains short, fixed label single-lead ECG windows (9–61 s). It is important to mention that the recordings there are 771 AFib records out of 8528 (~9% of the database), and 5154 records in normal rhythm class (~60%).

Another widely dataset that is widely used for model development is the MIT-BIH Atrial Fibrillation Database (AFDB), (Moody & Mark, 2001). It has long-term (10 hour long) two-lead, annotated Holter recordings. AFib episodes are intermittent in long duration records, and the class distribution is different in different patient records and depends on the window length and labeling method. Main drawback of this dataset is its small patient size (25 records).

Table 1. Publicly available databases, for AFib detection tasks.

Datasets	Records	Duration	Leads	Sampling rate (Hz)	Rhythm Labels
CinC2017 (Clifford et al., 2017)	8,528	9–61 s	1	300	AFib, N, other, noisy.
SHDB-AF (Tsutsui et al., 2025)	128	24 h	2	200	AFib, AFL, AT, PAT/NOD, N
AFDB (Moody & Mark, 1983)	25	10 h	2	250	AFib, AFL, J, N
Chapman–Shaoxing & Ningbo (Zheng et al., 2020)	10,646	10 s	12	500	11 rhythm classes
CinC2020 (Reyna et al., 2021)	43,101	Variable	12	Variable	27 rhythm classes
PTB-XL (Wagner et al., 2020)	21,799	10 s	12	500	71 classes
CPSC2018 (Liu et al., 2018)	6,877	6–60 s	12	500	9 rhythm classes

AFib – atrial fibrillation; AFL – atrial flutter; AT – atrial tachycardia; PAT/NOD – paroxysmal atrial tachycardia / nodal rhythm; J – junctional rhythm; N – normal rhythm or other non-AFib

Recently, a new Saitama heart database atrial fibrillation (SHDB-AF) has been introduced, that also includes long duration two-lead recordings, with detailed rhythm annotations (Tsutsui et al., 2025). Because SHDB-AF were acquired using clinical Holter monitoring, the recording conditions closely resemble other wearable device recording scenarios and some AFib detection pipelines have already utilized it (Ben-Moshe et al., 2023).

In addition, researchers also use private datasets and multi-lead clinical ECG databases for single-lead AFib studies that include large clinical 12-lead ECG recordings with multiple rhythm

annotations. When such multi-lead datasets are used for single-lead AFib classification, lead II is most selected to maintain consistency with wearable records (Liang et al., 2024). However, since these signals are recorded in clinical settings using standard ECG recording equipment, they are less representative of wearable ECG signals, when compared to SHDB-AF, AFDB.

It is important to mention that most of these datasets have been already preprocessed, and the signals are relatively clean. Therefore, the MIT-BIH Noise Stress Test Database (NSTDB) (Moody & Mark, 1984) is used as it has records of BW, MA and EM noise that is added to the ECG signal to simulate noise that is common in the wearable recordings.

1. 5. SIGNAL PREPROCESSING STRATEGIES

Signal processing is an important part of DL workflows, because DL models require inputs with consistent structure. Before the classification task, the ECG signal is commonly resampled to a target frequency, denoised, segmented, and normalized before model input. Z-score normalization is usually used to standardize signal amplitude (Liang et al., 2024).

One of the most critical steps in ECG signal processing is window length selection and segmentation, because it determines how much temporal information will be available for a classifier and determines how labels are assigned. In many AFib classification articles where already segmented databases (e.g. CinC2017) are used, truncation and zero-padding to fixed length are applied (Liang at al., 2024, Gao at al., 2021), because ECG records in these databases are relatively short and varies in length significantly (9 s to 61 s). While this makes the dataset convenient for benchmarking, it does not fully reflect the Holter monitor recordings that span over multiple hours and might have paroxysmal AFib episodes. However, when long-term ECG records are used (e.g. AFDB and other datasets), fixed-length windows are extracted by non-overlapping segmentation (Ben-Moshe et al., 2023; Li et al. 2026) or sliding window with overlap (Mousavi et al. 2020).

Regardless of segmentation strategy, there is no consensus on what segment length to use for AFib detection tasks. Reported window lengths range from short and less computationally expensive 5 s (Li et al., 2023; Mousavi et al., 2020), 10s (Li et al., 2025) 15 s segments (Li et al., 2024; Wu et al., 2025) to 30 s segments (Gao et al., 2021; Ben-Moshe et al., 2024). In some instances, window

length reaches up to 60 s (Liang et al., 2024), however, clear justification for specific length is rarely stated.

Labeling after segmentation is also important. In continuous Holter datasets with paroxysmal AF episodes, segment labels depend on an explicit rule because rhythm can change within a window. One approach labels a segment as AFib if any AFib occurs within that segment (Li et al., 2024). Most other authors use a majority rule for assigning segment labels when mixed rhythms occur within a window (Ben-Moshe et al., 2023; Mousavi et al., 2020). Some authors merge AFib and atrial flutter (AFL) into a single positive class (Li et al. 2024), However Biton et al. (2023) analysis showed that AFL was a major source of false negatives.

Overall, literature describes broad variation in segment length and labeling rules, with no clear consensus. Therefore, preprocessing remains a major source of variation across AFib classification studies.

1. 6. SIGNAL DENOISING STRATEGIES

Studies observed that AFib classifier accuracy drops with noise or motion artifacts are added to the ECG record (Li et al., 2026). Hence, signal denoising is an important step in AFib detection. One of the most used denoising strategies in AFib detection pipelines is bandpass filtering and wavelet-based denoising. Multiple authors use Butterworth high and low pass filters of 0.5-40 Hz (Zhang el al., 2019; Li et al., 2025). These filters are effective at removing baseline wander and powerline interference and preserving the morphology of the P-wave, QRS complex, and T-wave. Some authors also use Discrete wavelet transforms (DWT) with Daubechies wavelets are also sometimes used in 1D signal analysis due to their similarity to the ECG morphology (Li et al., 2024). Wavelet methods work by decomposing the signal, so that the noisy components to be reduced. However, wavelet based denoising needs careful parameter selection, otherwise they can distort the signal. Therefore, these methods are not sufficient if the signal is corrupted by complex noise, that is common in wearable devices.

In recent years, deep learning (DL) models are being applied to the ECG denoising task with promising results. DL-based approaches can learn complex, non-linear relationships between noise and clean ECG signals that traditional filtering methods cannot. A completely new and promising

direction in this field are the diffusion of probabilistic models (DDPMs). Diffusion models work by defining a forward process that gradually corrupts a clean signal x_0 by adding noise over T steps, producing increasingly noisy versions x_1, x_2, ..., x_t,. The model learns to reverse this process. It starts from a noisy observation, it iteratively removes noise step by step, conditioned on the corrupted input $\tilde{x}$, until a clean reconstruction is recovered. At each step, the network predicts the noise component present at that diffusion level and subtracts it. These models were originally developed for image data (Ho et al., 2020). Later, they were applied to the ECG signal denoising tasks (Hu et al., 2024; Han et al., 2025; Li et al., 2024). Diffusion models can have advantages over traditional methods because they may learn to remove complex noise that is a significant challenge in real-life ECG recordings (Li et al., 2025). However, how they would affect the downstream AFib classification tasks remains unclear.

1.7. ATRIAL FIBRILLATION DETECTION ALGORITHMS

Early automated AFib detection methods were handcrafted features for example R–R intervals, extracted from the ECG. These approaches are computationally light and can detect irregularity between R peaks well (Petrėnas et al. 2015). However, RR-only methods ignore the other complex morphological changes of the ECG that appear in AFib (e.g., the absence of the P wave), because they only use the timing between the R peaks for analysis. This can reduce their accuracy in practical applications, as AFib is also characterized by the lack of a P wave, and RR irregularity can be observed in subjects with other conditions. For these reasons, current research has increasingly shifted to automatic feature extraction.

One of the most used DL architectures AFib detection from raw ECG signals is 1D convolutional neural networks (CNNs). A CNN applies convolutional filters that slide over the signal to detect patterns in early layers and then composes these into higher-level representations in deeper layers (Goodfellow et al., 2016). This is useful for learning ECG morphological features (e.g., R peaks and P waves) directly from ECG (Xia et al. 2018). To increase model depth and avoid vanishing gradient issues, many authors include residual connections (Hannun et al., 2019; Xie et al., 2024). Residual blocks allow deeper layers to learn more abstract discriminative features. Also, because AFib detection depends on ECG patterns at different time scales, multiscale convolutions are useful (Wu et al. 2025). The main advantage of multiscale convolution is that it broadens the

receptive field. Small kernels extract local waveform features, while larger kernels extract broader rhythm patterns.

Another important development is the increasing use of signal representations beyond the 1D ECG. Some studies transform ECG signals into time–frequency representations and then apply 2D convolutional networks for classification (Li et al., 2023). The idea is that AFib can be detected by characteristic spectral structure. This also gave rise to two branch and multimodal models that jointly encode the raw signal and a spectral representation (Reich et al., 2023; Wu et al. 2025). The main advantage of time–frequency and multimodal approaches is that they can extract and combine additional information from different domains. This may improve robustness and classification accuracy, especially on short or noisy segments. However, this complicates preprocessing and increases memory and computation requirements.

Even though CNNs work well for local ECG beat morphology, they can be limited in analyzing the rhythm patterns of the longer signal. Therefore, many authors suggest using hybrid architecture that combines sequential architecture that can process data step by step and learn patterns over time. Research shows that CNN-Recurrent neural network (RNN) hybrid outperformed pure CNN (Zihlmann et al. 2017). Similarly, Petmezas et al. (2021) used CNN and long short-term memory (LSTM) to improve classification with focal loss to address class imbalance. And Biton et al. (2023) combined residual CNN-based embedding extraction with Gated recurrent unit (GRU) based sequence modeling. Their main strength is improved temporal modeling of longer signals.

More recently, Transformer-based models have been investigated. Liang et al., (2024) proposed CNN-Transformer hybrid, that outperforms CNN and CNN-LSTM baselines. The authors argue that CNNs are effective for local features but weak in modeling global relationships, while RNN-style solutions still suffer from degraded efficiency on long sequences. Transformer-based designs therefore represent a newer attempt to model rhythm context more comprehensively.

Table 2. Some specifications of reviewed single-lead AFib classification pipelines.

Architecture	**Datasets**	**Segments**	**Feature**	**Eval metric**	**Result**	**Article**
CNN+Transformer (CTRhythm)	CinC2017, CPSC2018, Chapman	60 s	Raw ECG	Acc, F1	85.4%, 83.1%	(Liang et al., 2024)

SC-CNN (2D CNN, Z-shaped)	AFDB, CinC2017, CPSC2018	5 s	Raw ECG (2D reconstructed)	Acc, Spe, Sen,	AFDB: 99.79, 99.72, 99.86 CinC2017: 95.51, 96.92, 94.06	(Li et al. 2023)
LCG-Net (lightweight dual-path CNN)	AFDB, SPHDB	15 s	Raw ECG	Acc, F1	98.39%, 98.38% (SPHDB)	(Li et al. 2024)
MFEG-Net (CNN+GRU, multiscale)	CinC2017, CPSC2018, AFDB	15 s	Raw ECG	Acc, F1	CinC2017: 93%, 88.3%	(Wu et al. 2025)
MGCNet (CNN+GRU+STFT, multimodal + contrastive learning)	AFDB, CPSC2021	10 s	Raw ECG + spectrogram	Acc, AUC, F1	95.1% 98.9% 93.3	(Li et al. 2025)
RTA-CNN + EN-Loss (residual temporal attention CNN)	CinC2017	30 s	Raw ECG	Acc	83.82%	(Gao et al. 2021)
BiRNN + hierarchical attention (HAN-ECG)	AFDB, CinC2017	5 s, 30-60 s	Raw ECG waves (75 pts/wave)	Se, Sp, Acc, AUC	AFDB: 99.08%, 98.54%, 98.81%, 99.86% CinC2017: 86.02%, 98.62%, 96.98%, 98.46%	(Mousavi et al. 2020)
RawECGNet (ResNet+DSU+BiGRU)	UVAF RBDB SHDB-AF	30 s	Raw ECG (single-lead)	F1	93% (SHDB-AF)	(Ben-Moshe et al. 2023)
CNN / CRNN ensemble	CinC2017	Variable	Raw ECG	F1	82%	(Zihlmann et al. 2017)

BiRNN + hierarchical attention (HAN-ECG)	AFDB, CinC2017	5 s, 30-60 s	Raw ECG waves (75 pts/wave)	Se, Sp, Acc	AFDB: 99.08%, 98.54%, 98.81%, CinC2017: 86.02%, 98.62%, 96.98%	(Mousavi et al. 2020)
CNN+ResNet+GRU (ArNet2)	UVAF, RBDB, SHDB, CPSC	60-beat RR windows	RR intervals	Se, Sp, F1	90 %; 96 %; 81 %	(Biton et al. 2023)

Acc – accuracy; AUC – area under the receiver operating characteristic curve; Pre – precision; pts – points; Se/Sen – sensitivity; Sp/Spe – specificity; STFT – short-time Fourier transform. s – seconds.

Overall, the literature shows shift from handcrafted toward automatic feature extraction from raw ECG and richer multimodal representations. CNNs with residual and multiscale designs remain strong baselines because they are effective at learning local morphological features. Transformer-based and multimodal time–frequency models appear promising because they integrate local morphology with broader temporal or spectral context.

1.8. LIMITATIONS OF CURRENT APPROACHES

Despite reported strong performance, current AFib detection workflows still have several limitations. Different papers use different datasets, and different evaluation scenarios, and sometimes external validation on unseen patient data is limited. In addition to this, segmentation procedure and resulting segment length and labeling strategy are not standardized. Across the reviewed literature, AFib models were trained on different segment lengths, but the rationale for using them is not explicitly explained or compared. Due to these differences in benchmarking, comparison between studies is difficult.

Also, many approaches still show limited robustness to noise and real-world recording conditions. Many models that try to test for noisy signals show that AFib detection accuracy deteriorates under noise (Li et al., 2026). In addition, denoising is often under investigated as an experimental factor. Many detection articles include filtering, but novel denoising methods are not introduced. This creates a research gap, where new diffusion-based denoising methods show promising results in current denoising research, but their effects on downstream AFib classification tasks are not evaluated in a controlled manner.

Overall, the current literature is limited by non-uniform benchmarking, inconsistent segment design, insufficient robustness analysis under realistic noise, and limited evaluation of denoising as part of the AFib classification pipeline.

METHODS

2.1 DATA

Two publicly available datasets were used for model training and evaluation. SHDB-AF (Tsutsui et al., 2025) dataset was selected for model training due to its large size. AFDB (Moody & Mark, 1983) was used for external cross-dataset evaluation as it is frequently used benchmark dataset in AFib detection studies. Both datasets include long-term Holter monitor records that have heart rhythm annotations. For denoising and noise robustness tasks, NSTDB (Moody et al., 1984) dataset was utilized to simulate three types of noise (BW, MA, EM) common in wearable device recordings. The main characteristics of all three data sets used are summarized in Table 3.

Table 3. Main characteristics of the datasets used in this thesis

Dataset	**Subjects**	**Annotated records**	**Recording length**	**Leads**	**Sampling rate**	**Classes / rhythm labels**
SHDB-AF	122	98	24 h	2	200 Hz	AFib, AFL, AT, PAT/NOD, N
MIT-BIH AFDB	25	23	10 h	2	250 Hz	AFib, AFL, J, N
MIT-BIH NSTDB	–	3 noise records	30 min	–	–	BW, MA, EM noise; SNR levels from 24 dB to -6 dB

AFib – atrial fibrillation; AFL – atrial flutter; AT – atrial tachycardia; PAT/NOD – paroxysmal atrial tachycardia / nodal rhythm; J – junctional rhythm; N – normal rhythm or other non-AFib rhythm; BW – baseline wander; MA – muscle artifact; EM – electrode motion artifact; SNR – signal-to-noise ratio; dB – decibels.

2.2 DATA PREPARATION

Before segmentation into fixed-length windows, ECG records from both datasets were pre-processed. A 4th-order Butterworth band-pass filter with cut-off frequencies of 0.5–40 Hz was used to reduce baseline wander and high-frequency noise. Also, AFDB signals were down sampled

to 200 Hz to match the sampling frequency of SHDB-AF. Finally, records were divided into non-overlapping fixed-length windows, and each window was assigned a class label using majority rule. Positive class (AFib = 1) was assigned if at least 50 % of its samples had the AFib rhythm label. This threshold-based rule is consistent with the approach used in related work (Moshe et al., 2023). Finally, resulting windows were normalized using z-score normalization.

Dataset splits followed an interpatient paradigm, so that all windows from the same patient appeared in only one partition. From SHDB-AF 20% of patients were held out test set and the remaining 80% formed the train and validation section. And the whole AFDB dataset was used for cross dataset testing. For SHDB-AF records were additionally stratified by the percentage of AF-labelled windows per patient, so each partition contained a similar percentage of AFib windows. Patient records were assigned to three bins: no AFib, low AFib burden (≤ 20%), and high AFib burden (> 20%). This stratification is used in related work (Li et al., 2026), because it ensures that the AFib class proportion remains consistent across partitions.

For noise robustness experiments all clean test windows were contaminated with three types of noise using a segment-adaptive protocol (Šimėnas et al., 2025). Each clean ECG window x_i was corrupted by an NSTDB noise segment $n_i$of equal length according to:

$$\tilde{x}_i = x_i + \alpha_i n_i, \tag{1}$$

where the scaling factor $\alpha_i$was defined as:

$$\alpha_i = \frac{\rho_i}{R_n/(R_x+\epsilon)+\epsilon}, \rho_i \sim U(0.2,2.0), \tag{2}$$

where R_x and R_n denote the dynamic ranges of the ECG and noise segments, ρ_i is a random factor between 0.2 and 2.0 that controls contamination intensity, and $\epsilon = 10^{-12}$ prevents numerical issues. This created different levels of noise across windows, which is closer to real-world noise than using a fixed noise level.

2.3 DEEP NEURAL NETWORKS

Three DL classification models were selected to test the different approaches to AFib detection from single-lead ECG signals that are relevant to objectives of this thesis. The main goal was to

determine which model could reach highest evaluation metrics in this task. In addition, a diffusion-based denoising model was used to assess whether upstream signal denoising can recover classification performance degraded by noise.

Convolutional Neural Network–Transformer Rhythm Classifier (CTRhythm) is a hybrid DNN that integrates a 1D CNN with a Transformer encoder (Liang et al., 2024). The reasoning for this design selection is that CNNs are suited for extraction of local morphological features of the ECG, and Transformer encoder can learn rhythm features across the full signal window in parallel. This helps the model overcome the limited receptive field of pure CNNs and vanishing gradient issue that comes with sequential architectures. (Fig. 4).

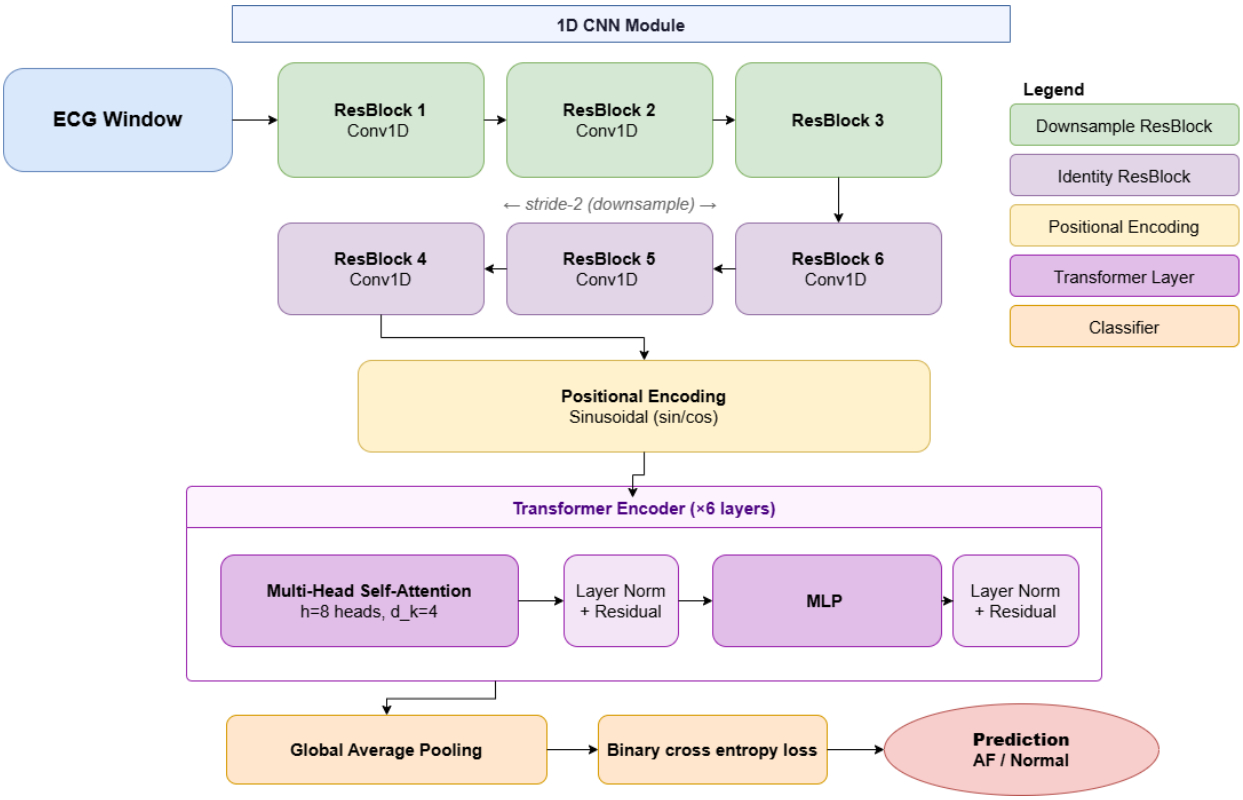

Figure 4. Architecture of the CTRhythm model (Liang et al., 2024).

CTRhythm architecture includes six residual CNN blocks followed by positional encoding and six Transformer encoder layers with eight-head self-attention module. Transformer encoder increased F1 score to 0.831 and outperformed purely CNN model (macro F1 was 3.2% lower) and CNN-LSTM hybrid (2.3% lower macro F1 score).

Multiscale feature-enhanced gating network (MFEGNet) (Wu et al., 2025) was proposed to address limited feature extraction and reduced performance in noisy signals. The architecture (Fig. 5) includes a multiscale convolution module, two feature enhancement blocks, three residual blocks that perform down sampling for deep feature extraction, and a GRU layer that analyses patterns in heart rhythm.

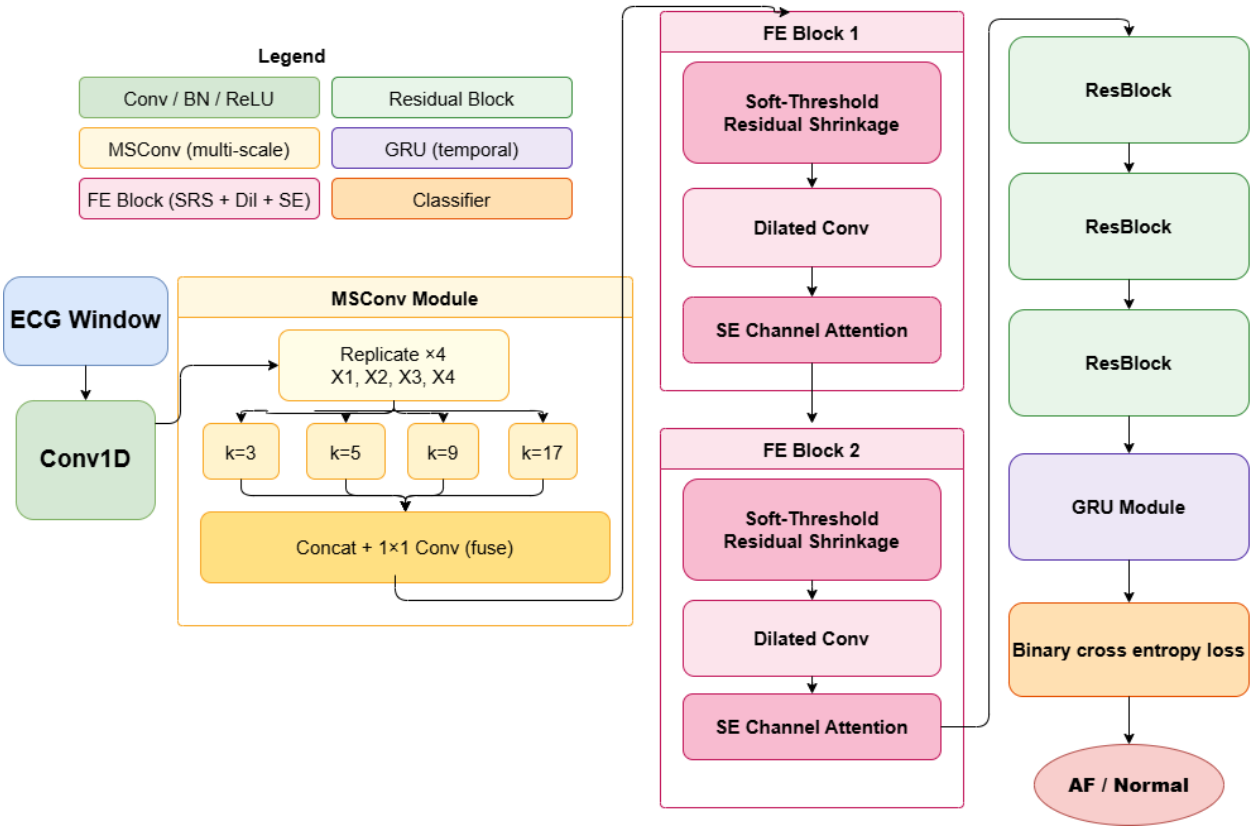

Figure 5. Architecture of the MFEG Net model (Wu et al., 2025).

The multiscale convolution module applies four parallel convolutional branches that allow the model to extract ECG morphological features at multiple temporal scales simultaneously. To suppress noisy and redundant features each feature enhancement block combines a soft-threshold residual shrinkage unit, a dilated convolution to widen the receptive field, and a Squeeze-and-Excitation module that assigns a learned weight to each feature channel. Therefore, channels that have AFib-relevant information are amplified and others are suppressed.

Both CTRhythm and MFEGNet were trained for multiclass rhythm classification in the original articles. For this study, architecture was adapted for binary AFib detection by replacing the loss function with a binary version of cross-entropy loss (BCE).

Because AFib can be detected from both spectrograms and raw signals, multimodal gated contrastive network (MGC) was proposed to improve AFib detection robustness and cross-dataset generalization using dual-branch design (Li et al., 2026). The time-domain branch processes the raw ECG segment through a 1D convolutional encoder, and the frequency-domain branch converts the same segment into a Short-Time Fourier Transform spectrogram and processes it through a 2D convolutional encoder (Fig. 6).

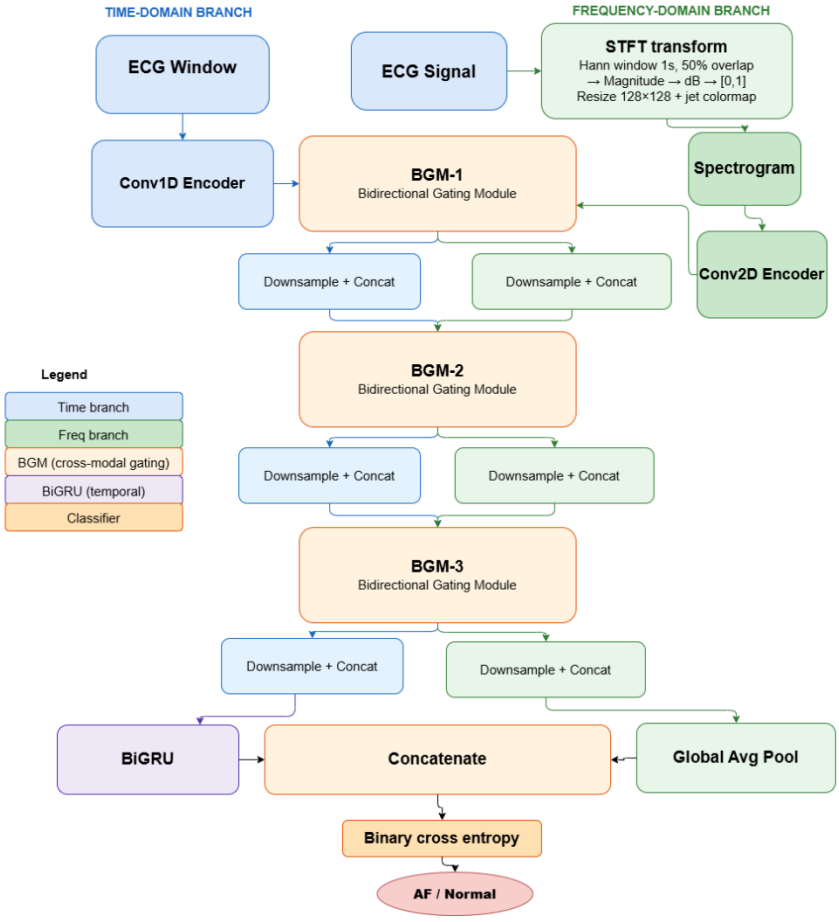

Figure 6. Architecture of the MGCNet model (Li et al., 2026).

The reasoning behind this is that AFib manifests as irregular rhythm intervals in the time domain and chaotic atrial activity in the frequency domain. Therefore, processing simultaneously provides a more complete representation than either domain alone.

At three hierarchical levels, the branches are connected through a Bidirectional Gating Module, which computes per-channel weight vectors and applies them to suppress or amplify channels in the opposite branch. The time-domain branch uses BiGRU to model temporal patterns bidirectionally, while the frequency branch uses global average pooling to summarize spectral content.

Also, a pretrained conditional score-based diffusion model (DeScoD-ECG) (Li et al., 2025), was used to evaluate if deep learning-based denoising can improve classification performance degraded by noise.

2.4. EXPERIMENTAL DESIGN

The experiments were designed so that all classification models and impact of denoising would be systematically investigated. The same data splits and evaluation metrics were applied across all experiments. All computations were performed on the VU MIF ITAPC supercomputer, which provides 1912 CPUs, 17 TB of RAM, 32 GPUs, and 0.5 PB of disk storage. The GPU cluster includes three NVIDIA DGX-1 systems and two IBM Power9 AC922 systems equipped with NVIDIA Tesla V100 GPUs with 32 GB memory.

During the first stage of experiments selected DL models were evaluated using the same window length to identify which architecture generalizes most reliably on unseen patient data. All models were trained on the SHDB-AF dataset with 10-second input windows and BCE. The best performing model was evaluated on the held-out SHDB-AF test set, and the external AFDB dataset. One model was selected for further experiments.

Also, to gain insight into the features learned by the best performing model, the attention weights of the Transformer encoder layers were visualized using the Attention Rollout technique used by the original the CTRythm article (Liang et al., 2024). A few examples were selected from the test

set, and their attention maps were inspected to assess whether the regions highlighted by the model are clinically meaningful.

The second stage of experiments were performed to investigate the effect of input window length and loss function on AFib detection performance. First, the best performing model was trained and evaluated on four window lengths (10, 15, and 30 seconds) used in the related literature. The window length that produced the best results was selected for loss function selection experiments. Because AFib windows make up 20% of the training dataset, model was retrained using focal loss to see if it can improve model performance over the BCE loss function used in the original model:

$$\mathrm{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t), \quad (3)$$

where $(1 - p_t)^x$ is the modulating factor that reduces the loss contribution of easy examples, γ is the focusing parameter that controls the rate of down-weighting (used $\gamma = 2.0$), and α_t is a class-weighting factor that addressed class imbalance (used $\alpha = 0.8$). Focal loss was selected as being shown to be effective for imbalanced ECG classification tasks (Petmezas et al., 2021).

Finally, our aim was to simulate real world ECG recordings when models must classify noisy ECG windows and investigate degradation of feature extraction quality of different classifiers and determine if denoising would improve classification performance. For this purpose, additional noise from the MIT-BIH Noise Stress Test Database (NSTDB) was added to subsection of test sets, and model performance was re-evaluated. The NSTDB noise injection procedure described in section 2.2 was applied to the test windows and the noisy signals were denoised using a pre-trained diffusion-based denoising model, DeScoD-ECG (Šimenas et al., 2025). Due to the computational costs of denoising algorithm, only a small subsection of test signals had the noise added and later denoised. Both noisy and denoised signals were evaluated using the selected classifier to see how the denoising algorithm affected model performance.

2.5. EVALUATION

AFib detection is framed as a binary classification problem. Each ECG window is assigned either the positive class (AFib = 1) or the negative class (N = 0). Classification outcomes are organized into a confusion matrix as shown in Table 4.

Table 4. Confusion matrix for binary AFib versus non-AFib classification.

	Predicted AFib	Predicted non-AFib
True AFib	TP (true positive)	FN (false negative)
True non-AFib	FP (false positive)	TN (true negative)

AFib – atrial fibrillation;

A true positive (TP) represents an AFib window correctly predicted as AFib. A true negative (TN) is a non-AF window correctly predicted as non-AF. A false positive (FP) is a non-AFib window incorrectly predicted as AF, and a false negative (FN) is an AFib window incorrectly predicted as non-AF.

Recall measures the ability to correctly identify AFib windows:

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

Precision measures the proportion of AFib predictions that are correct:

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

F1 score is the harmonic mean of recall and precision:

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

In addition to the class-specific F1 score, the macro F1 score is computed as the unweighted mean of the F1 scores obtained for each class:

$$Macro\ F1 = \frac{F1_{AF} + F1_{N}}{2}$$

EXPERIMENTAL RESULTS

3.1. MODEL SELECTION

CTRhythm, MFEGNet, and MGCNet classifiers selected during the literature review were trained on the SHDB-AF dataset and evaluated on held out test set of the same dataset consisting of unseen patient records and cross-dataset tested on AFDB dataset using the same window length. The aim was to evaluate how trained models perform on unseen data and to identify which model provides the most consistent and clinically reliable AFib detection. Results are reported in Table 5 (Vektarytė et al., 2026).

Table 5. Comparison of atrial fibrillation detection performance across three classifiers on 10-second windows.

Model	Test set	Macro F1	AFib F1	Recall	Precision
CTRhythm	SHDB-AF	**0.906**	0.856	**0.923**	0.802
	AFDB	**0.865**	0.827	0.728	**0.956**
MGCNet	SHDB-AF	0.902	**0.881**	0.852	**0.910**
	AFDB	0.860	**0.839**	**0.810**	0.869
MFEGNet	SHDB-AF	0.867	0.832	0.801	0.890
	AFDB	0.787	0.720	0.616	0.867

The highest values are highlighted in bold.

On the external AFDB dataset CTRhythm obtained the highest macro F1 score (0.865) and the highest AFib precision (0.956). MGCNet achieved similar macro F1 score (0.860), but detected a larger proportion of true AFib windows, as it had highest AFib recall (0.810), which was 11.3% higher than CTRhythm (0.728). Both models achieved better performance when evaluated on held out SHDB-AF test set. Macro F1 was 4.7% and 4.9% higher for CTRhythm and MGCNet respectively compared to external evaluation. However, during external evaluation MFEGNet showed the largest drop in performance, as macro F1 decreased from 0.867 to 0.787, and the model demonstrated the weakest performance across nearly all evaluated metrics on the external dataset. CTRhythm was selected for further experiments because it achieved better classification performance values with the lowest cross-fold variance. The macro F1 score was 9.9% and 0.6% higher than those of MFEGNet and MGCNet respectively on the external dataset.

To gain insight into CTRhythm's classification decisions, attention weights from the Transformer encoder were extracted and visualized over representative ECG segments for an AFib window and non-AFib windows (Fig. 7).

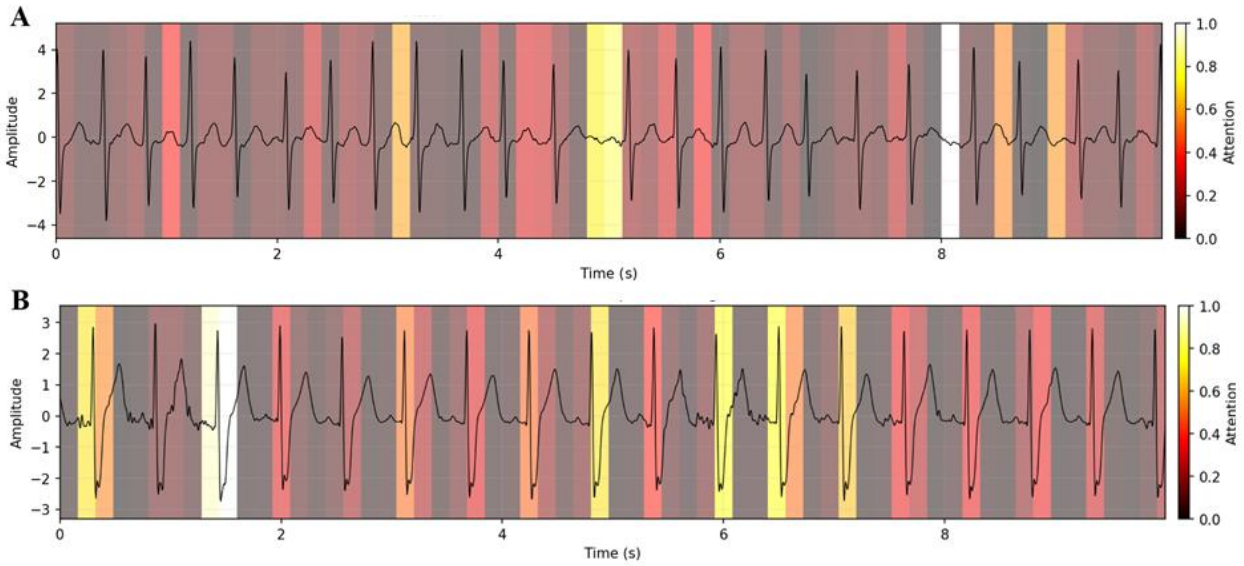

Figure 7. Visualisation of CTRhythm Transformer attention weights for AFib (A) and non-AFib windows (B). Lighter colors indicate regions that had greater influence in classification decision.

In the AFib segment, attention was focused on the P wave region and in non-AFib window attention shows higher weights concentrated around the QRS complexes. This suggests that CTRhythm extracted clinically relevant features of the AFib and non-AFib windows. However, these attention patterns were not fully consistent across all examined windows, because of variability of ECG signals across patients.

3.2. WINDOW LENGTH SELECTION

Further experiments investigated the effect of ECG window length on AFib detection performance for CTRhythm classifier. The same experimental protocol was followed, and performance was evaluated across 10, 15, and 30 seconds window lengths.

As seen from Table 6, CTRythm recall increased with longer windows (0.923, 0.947, and 0.953 for 10, 15, and 30 seconds), but precision lowered (0.802, 0.661, and 0.638 respectively). As the model detected more of true AFib episodes but also made more false positive predictions.

Table 6. Comparison of AFib detection performance for CTRhythm across different window lengths.

Window lenght	Test set	Macro F1	AFib F1	Recall	Precision
10s	SHDB-AF	**0.906**	**0.856**	0.923	**0.802**
	AFDB	**0.865**	**0.827**	**0.728**	**0.956**
15s	SHDB-AF	0.863	0.778	**0.947**	0.661
	AFDB	0.811	0.758	0.643	0.922
30s	SHDB-AF	0.848	0.763	0.953	0.638
	AFDB	0.823	0.771	0.697	0.885

The highest values are highlighted in bold; s – seconds.

However, on the external AFDB dataset, longer windows did not consistently improve performance. The 10-second window achieved stronger generalization, as macro F1 (0.865) was 6.7% and 5.1% higher than the 15-second (0.811) and 30-second (0.823) windows respectively. It also showed highest AFib F1 (0.827), recall (0.728) and precision (0.956). These results suggest that shorter windows generalize better across datasets, therefore 10 second windows were used for further experiments.

3.3. LOSS FUNCTION SELECTION

To evaluate if model performance could be improved by addressing class imbalance in training dataset, focal loss ($\alpha = 0.8$, $\gamma = 2.0$) was compared against standard BCE loss.

As seen from Table 7, on the external AFDB dataset, focal loss consistently improved CTRhythm performance. It achieved a higher macro F1 (0.875), which was 1.2% higher than BCE, a higher AFib F1 (0.840, 2.1% higher), and a higher recall (0.776, 4.8% higher). However, it also increased the amount of false positive AFib predictions as precision was 3.5% lower when compared to standard BCE.

Table 7. Comparison of atrial fibrillation detection performance for CTRhythm trained with binary cross-entropy and focal loss on 10-second windows.

Loss function	Test set	Macro F1	AFib F1	Recall	Precision
BCE	SHDB-AF	**0.906**	**0.856**	0.923	**0.802**
	AFDB	0.865	0.827	0.728	**0.956**
Focal	SHDB-AF	0.898	0.848	**0.936**	0.775
	AFDB	**0.875**	**0.840**	**0.776**	0.921

The highest values are highlighted in bold. BCE – binary cross entropy.

3.5. DENOISED SIGNAL CLASSIFICATION

To check model noise robustness, the model was re-evaluated on the same test sets contaminated with 3 different noise types (BW, MA, EM), using the protocol described in Section 2.2, and were denoised by a pretrained diffusion-based denoiser DeScoD-ECG (Šimėnas et al., 2025) prior to classification (Fig. 8). Model performance was then compared against the noisy baseline under identical conditions (Table 8).

Noise contamination degraded performance across all noise types and both datasets, with AFib recall being the most severely affected metric in all conditions. On SHDB-AF, BW noise caused the largest recall to drop from 0.982 to 0.250 (74.5% decrease), EM caused it to drop from 0.982 to 0.304, (69% decrease) and MA caused 67.3% decrease (from 0.982 to 0.321). On AFDB, degradation was even more severe as recall decreased by 84.4%, 93.4%, and 89.0% for BW, EM and MA respectively, as the model missed majority of true AFib episodes.

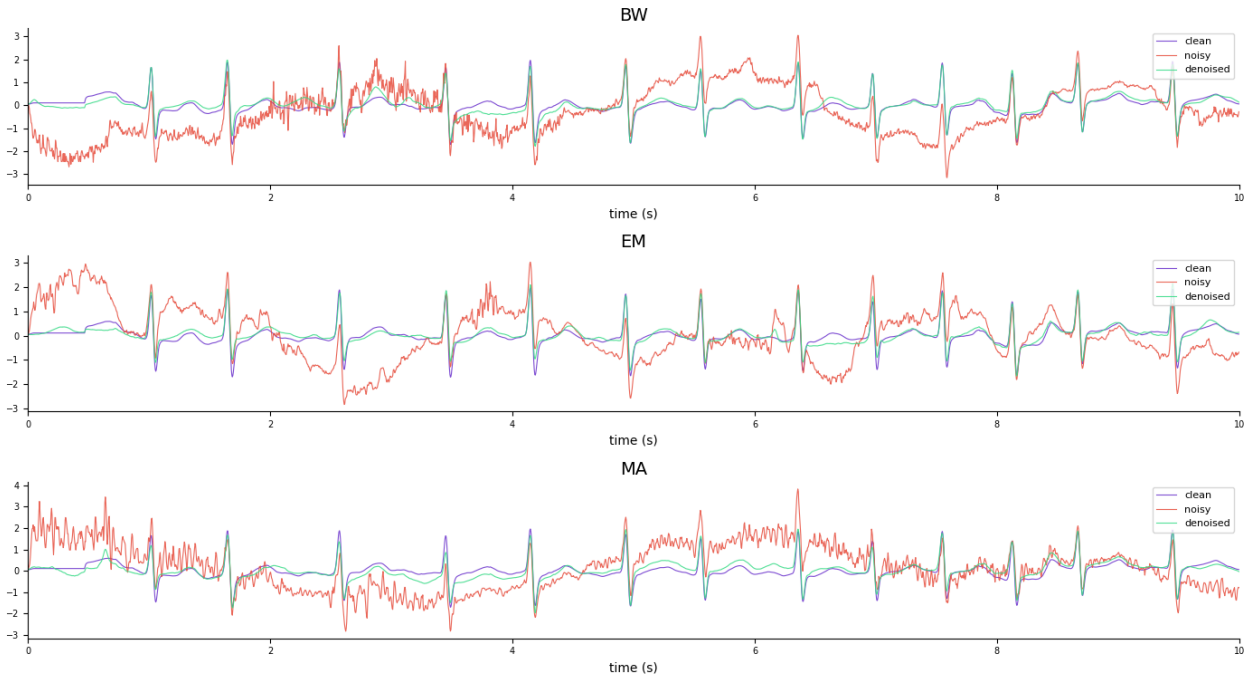

Figure 8. Visualization of the ECG windows with baseline wander (BW), electrode motion (EM) and motion artifact (MA) noise added. Clean signals shown in violet, noisy in orange and denoised in green colour.

Denoising improved model performance, however it was dependent on noise type and dataset. Signals corrupted with BW noise could be denoised with the most consistent results. The most complete recovery was observed under BW noise on SHDB-AF, as recall was restored from 0.250 to 0.982 in SHDB-AF dataset. However, restoration of signals corrupted with MA and EM was more limited as recall reached only 0.732 and 0.518 respectively which is below the clean baseline of 0.982. On AFDB, signal recovery was even less complete, as macro F1 score reached 0.715 (BW), 0.592 (MA), and 0.663 (EM), which were lower when compared to the clean signal baseline of 0.814. Across all conditions, precision recovered more completely than recall, which suggests that the model tends to miss more AFib episodes during classification.

Table 8. AFib detection performance of CTRhythm on noise-contaminated and denoised signals across three noise types.

Noise type	Signal	Test set	Macro F1	AFib F1	Recall	Precision
None	Clean	SHDB-AF	0.977	0.974	0.982	0.965
	Clean	AFDB	0.814	0.769	0.634	0.978
BW	Noisy	SHDB-AF	0.561	0.373	0.250	0.737
	Denoised	SHDB-AF	**0.954**	**0.948**	**0.982**	**0.917**
	Noisy	AFDB	0.407	0.157	0.099	0.389
	Denoised	AFDB	**0.715**	**0.630**	**0.479**	**0.919**
EM	Noisy	SHDB-AF	0.605	0.442	0.304	0.810
	Denoised	SHDB-AF	0.715	0.630	0.518	0.806
	Noisy	AFDB	0.368	0.073	0.042	0.273
	Denoised	AFDB	0.663	0.549	0.394	0.903
MA	Noisy	SHDB-AF	0.642	0.487	0.321	1.000
	Denoised	SHDB-AF	0.832	0.796	0.732	0.872
	Noisy	AFDB	0.349	0.104	0.070	0.200
	Denoised	AFDB	0.592	0.430	0.282	0.909

The highest values are highlighted in bold; BW – baseline wander, MA – muscle artifact, EM – electrode motion noise.

DISCUSSION

The results of this study demonstrate that the choice of DL architecture and signal preprocessing strategies influence AFib detection performance. It also shows the importance of addressing class imbalance and noise that is present in clinical ECG recordings.

All architectures were hybrid CNN-based models, that additionally included sequential (GRU and BiGRU) or transformer modules. This design is consistent with newest AFib detection studies, where CNN-based hybrid models dominate. Pure CNNs are used for extracting beat-level features, and sequential or Transformer blocks improve model performance when it comes to learning rhythm patterns over longer signal windows. In this experimental study, CTRhythm archieved the highest performance metrics on clean signals (macro F1 - 0.865) which reflects Transformers' ability to learn global rhythm patterns better then sequential models due to its attention mechanism being able to compare different points across the whole window, while sequential modules used by other models may introduce the risk of losing information from earlier parts of the signal. These results are supported by the original CTRhythm article (Liang et al. 2024), that showed that removing the Transformer encoder reduced macro F1 score from 0.831 to 0.799 and replacing it with LSTM layers reduced performance to 0.808. Despite a different approach, MGCNet demonstrated that an additional frequency branch can also reach high classification performance with macro F1 score 0.860. However, computing spectrograms can also increase computational costs during training and may be less suitable for wearable device deployment.

MFEGNet's archieved lower performance on clean signals (macro F1 - 0.787) which can be explained by different training protocols used by the original authors. Both CTRhythm and MFEGNet were originally trained on CinC2017 dataset that includes short windows with labels assigned at the record level and does not require segmentation. Holter monitoring datasets used in this experiment require fixed-window segmentation where labels are assigned using the majority rule, which can increase the risk of misclassification when windows contain mixed rhythms.

The window length and loss function selection experiment showed that 10-second windows generalize better on the external dataset, likely because this setup leads to more training samples and contain fewer ambiguously labelled windows. This finding is consistent with the current cardiology guidelines that highlight that previously used 30 second requirement for AFib detection might miss shorter AFib windows and is not recommended in clinical diagnosis (Van Gelder et

al., 2024). Also, focal loss resulted in a 4.8% increase in AFib recall on the external dataset, with 3.5% decrease in precision over BCE. In AFib detection tasks this is an acceptable trade-off because recall is more important in clinical settings. This is because a missed AFib episode has a greater risk for the patient than a false positive. Undetected and untreated AFib can lead to deadly outcomes (Masiliūnas et al.), and a false possitive could be ruled out by a medical professional.

However, model performance degraded below clinically acceptable metrics, when noise was simulated on the test set. On AFDB dataset, AFib recall collapsed most severely under BW and MA, dropping to 0.099 and 0.070 respectively, and macro F1 fell to 0.407 and 0.349. EM noise and MA are especially difficult to remove using standard preprocessing techniques due to the fact that they overlap with the frequency of the ECG signal. Therefore, a pretrained DeScoD-ECG denoiser was used as a preprocessing step. Diffusion-based models can learn to reconstruct the underlying clean signal during the reverse diffusion process and have shown promising results (Li et al., 2024a) but have not been evaluated during AFib classification tasks.

DeScoD-ECG was able to recover classification performance for BW, with macro F1 rising from 0.561 to 0.954 on SHDB-AF. For MA and EM noise, signal recovery was more limited. On the same dataset, denoised macro F1 reached only 0.832 and 0.715 respectively. In both cases the model had fewer false positives (precision: 0.872 for MA, 0.806 for EM), but it still missed a lot of AFib windows as the recall remained lower (0.732 for MA and 0.518 for EM) when compared to model performance on noisy signals. This could be because the denoiser introduced artifacts that impacted the classification accuracy and did not successfully restore the AFib-specific features needed for accurate classification, because it was pretrained on a different, non-AFib dataset. In the future, classification performance could be further improved by the integration of a signal quality assessment algorithm prior to classification (Ben-Moshe et al., 2023), where windows that fall below an experimentally determined quality threshold would be discarded both during training and evaluation.

Several limitations of this study limit its conclusions. Noise was artificially simulated by adding NSTDB recordings to clean signals, which might not fully represent noise found in real Holter monitoring. Also, this study evaluated per-window classification accuracy rather than episode-level AFib burden estimation (Ben-Moshe et al., 2023) which could be more clinically useful for Holter monitoring.

CONCLUSIONS

1. CTRhythm achieved macro F1 score of 0.865 on clean windows of the external AFDB dataset, which was 9.9% higher than MFEGNet and 0.6% higher than MGCNet macro F1 values.

2. CTRhythm archieved the highest macro F1 score using a 10-second window length on AFDB datset, which was 6.7% higher than the 15-second and 5.1% higher than the 30-second windows.

3. Focal loss increased AFib recall by 4.8% (from 0.728 to 0.776) and decreased precision by 3.5% (from 0.956 to 0.921) on cross dataset evaluation.

4. Noise contamination caused macro F1 score to drop from 0.814 to 0.407, 0.349, and 0.368 under BW, MA, and EM noise respectively on the external dataset, while AFib recall decreased by 84.4%, 89.0%, and 93.4% across the three noise types respectively.

5. Diffusion-based denoising recovered performance most effectively under baseline wander, as it restored macro F1 value on SHDB-AF from 0.561 to 0.954, but achieved more limited recovery under MA and EM noise on AFDB dataset.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor dr. Jolita Bernatavičienė who spent a lot of time advising and sharing her knowledge with me. I also want to thank Simonas Šimėnas, Jonas Mindaugas Rimšelis, Zigmantas Kęstutis Juškevičius and other researchers that work in the Institute of Data Science and Digital Technologies for their help and support.

I am also thankful for the high-performance computing resources provided by the Information Technology Open Access Center of Vilnius University.

This research has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-ITP-25-9.

During the writing of this thesis, AI-based language assistance tools: Grammarly and Claude were used to improve grammar, spelling, sentence structure, and readability in some sections. The use of these tools was limited to linguistic editing. The research design, data processing, analysis, interpretation, and conclusions are entirely my own work.

REFERENCES

1. Ben-Moshe, N., Tsutsui, K., Biton, S., Sörnmo, L. and Behar, J.A. (2023) 'RawECGNet: Deep learning generalization for atrial fibrillation detection from the raw ECG', arXiv preprint arXiv:2401.05411. https://doi.org/10.48550/arXiv.2401.05411
2. Biton, S., Aldhafeeri, M., Marcusohn, E., Tsutsui, K., Szwagier, T., Elias, A., Oster, J., Sellal, J.M., Suleiman, M. and Behar, J.A. (2023) 'Generalizable and robust deep learning algorithm for atrial fibrillation diagnosis across geography, ages and sexes', npj Digital Medicine, 6(1), 44. https://doi.org/10.1038/s41746-023-00791-1
3. Chiang, C.-E., Naditch-Brûlé, L., Murin, J., Goethals, M., Inoue, H., O'Neill, J., Silva-Cardoso, J. and Steg, P.G. (2012) 'Distribution and risk profile of paroxysmal, persistent, and permanent atrial fibrillation in routine clinical practice: Insight from the real-life global survey evaluating patients with atrial fibrillation international registry', Circulation: Arrhythmia and Electrophysiology, 5(4). https://doi.org/10.1161/CIRCEP.112.970749
4. Clifford, G.D., Liu, C., Moody, B., Lehman, L.-W.H., Silva, I., Li, Q., Johnson, A.E. and Mark, R.G. (2017) 'AF classification from a short single lead ECG recording: The PhysioNet/Computing in Cardiology Challenge 2017', Computing in Cardiology, 44. https://doi.org/10.22489/CinC.2017.065-469
5. Gao, Y., Wang, H. and Liu, Z. (2021) 'An end-to-end atrial fibrillation detection by a novel residual-based temporal attention convolutional neural network with exponential nonlinearity loss', Knowledge-Based Systems, 212, 106589. https://doi.org/10.1016/j.knosys.2020.106589
6. Goldberger, A.L., Goldberger, Z.D. and Shvilkin, A. (2017) Goldberger's Clinical Electrocardiography: A Simplified Approach. 9th edn. Philadelphia, PA: Elsevier.
7. Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. Cambridge, MA: MIT Press.
8. Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P. and Ng, A.Y. (2019) 'Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network', Nature Medicine, 25(1), pp. 65–69. https://doi.org/10.1038/s41591-018-0268-3

9. Ho, J., Jain, A. and Abbeel, P. (2020) 'Denoising diffusion probabilistic models', in Advances in Neural Information Processing Systems (NeurIPS). https://doi.org/10.48550/arXiv.2006.11239

10. Hu, Z.-D., Hong, Y., Huang, J.-Y., Chen, K.-H., Zhao, W.-Q., Grau, A., Guerra, E., Wang, C.-S. and Zhang, F.-Q. (2024) 'DMAM-ECG: A diffusion model with self-attention module for ECG signal denoising', Journal of Network Intelligence, 9(3), pp. 1278–1296.

11. Han, T.-S., Heo, J.-W., Kim, H., Lee, C.-H., Huh, H., Choi, E.-K., & Kim, D.-J. (2025). Diffusion-based electrocardiography noise quantification via anomaly detection. arXiv. https://doi.org/10.48550/arXiv.2506.11815

12. Ioannidis, P., Zografos, T., Christoforatou, E., Kouvelas, K., Tsoumeleas, A. and Vassilopoulos, C. (2021) 'The electrophysiology of atrial fibrillation: From basic mechanisms to catheter ablation', Cardiology Research and Practice, 2021, 4109269. https://doi.org/10.1155/2021/4109269

13. Li, H., Ditzler, G., Roveda, J. and Li, A. (2024a) 'DeScoD-ECG: Deep score-based diffusion model for ECG baseline wander and noise removal', IEEE Journal of Biomedical and Health Informatics, 28(9), pp. 5081–5092. https://doi.org/10.1109/JBHI.2023.3237712

14. Li, H., Wei, J., Xiao, J., Lai, Y., Liu, M., Lv, S. and Ouyang, X. (2026) 'Robust and generalizable atrial fibrillation detection from ECG using time-frequency fusion and supervised contrastive learning', arXiv preprint arXiv:2601.10202. https://arxiv.org/abs/2601.10202

15. Li, Y., Chen, M., Jiang, X. and Liu, L. (2024b) 'LCG-Net: An atrial fibrillation detection algorithm based on lightweight design architecture and feature fusion strategy', Biomedical Signal Processing and Control, 91, 106016. https://doi.org/10.1016/j.bspc.2024.106016

16. Li, Y., Zhang, L., Zhu, L., Liu, L., Han, B., Zhang, Y. and Wei, S. (2023) 'Diagnosis of atrial fibrillation using self-complementary attentional convolutional neural network', Computer Methods and Programs in Biomedicine, 238, 107565. https://doi.org/10.1016/j.cmpb.2023.107565

17. Liang, Z., Yang, C., Yu, Z., Fu, Y., Ren, B., Lin, M., Li, Q., Liu, X., Chen, Y. and Xia, L.C. (2024) 'CTRhythm: Accurate atrial fibrillation detection from single-lead ECG by

convolutional neural network and transformer integration', medRxiv. https://doi.org/10.1101/2024.10.26.24316175

18. Lindström, M., DeCleene, N., Dorsey, H. et al. (2022) 'Global burden of cardiovascular diseases and risks collaboration, 1990–2021', Journal of the American College of Cardiology, 80(25), pp. 2372–2425. https://doi.org/10.1016/j.jacc.2022.11.001

19. Liu, C., Zhao, L., Wei, S., He, Z., Li, J. and Ng, E.Y.K. (2018) 'An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection', Journal of Medical Imaging and Health Informatics, 8(7), pp. 1368–1373. https://doi.org/10.1166/jmihi.2018.2442

20. Llinares, R. and Igual, J. (2011) 'Exploiting periodicity to extract the atrial activity in atrial arrhythmias', EURASIP Journal on Advances in Signal Processing, 2011, 134. https://doi.org/10.1186/1687-6180-2011-134

21. Masiliūnas, R., Dapkutė, A., Grigaitė, J. et al. (2022) 'High prevalence of atrial fibrillation in a Lithuanian stroke patient cohort', Medicina, 58(6), 800. https://doi.org/10.3390/medicina58060800

22. Moody, G.B. and Mark, R.G. (1983) 'A new method for detecting atrial fibrillation using R–R intervals', in Computers in Cardiology, 10, pp. 227–230.

23. Moody, G.B., Muldrow, W.E. and Mark, R.G. (1984) 'A noise stress test for arrhythmia detectors', in Computers in Cardiology, 11, pp. 381–384.

24. Mousavi, S., Afghah, F. and Acharya, U.R. (2020) 'HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks', Computers in Biology and Medicine, 127, 104057. https://doi.org/10.1016/j.compbiomed.2020.104057

25. Petrėnas, A., Marozas, V. and Sörnmo, L. (2015) 'Low-complexity detection of atrial fibrillation in continuous long-term monitoring', Computers in Biology and Medicine, 65, pp. 184–191. https://doi.org/10.1016/j.compbiomed.2015.01.019

26. Petmezas, G., Haris, K., Stefanopoulos, L., Kilintzis, V., Tzavelis, A., Rogers, J.A., Katsaggelos, A.K. and Maglaveras, N. (2021) 'Automated atrial fibrillation detection using a hybrid CNN–LSTM network on imbalanced ECG datasets', Biomedical Signal Processing and Control, 63, 102194. https://doi.org/10.1016/j.bspc.2020.102194

27. Reich, C., Rohr, M., Kircher, T. and Hoog Antink, C. (2023) 'On the atrial fibrillation detection performance of ECG-DualNet', in 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). https://doi.org/10.1101/2023.06.01.23290822

28. Reyna, M.A., Perez Alday, E.A., Gu, A., Liu, C., Seyedi, S., Bahrami Rad, A. and Clifford, G.D. (2021) 'Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020', Physiological Measurement, 41(12), 124003. https://doi.org/10.1088/1361-6579/abc960

29. Sibomana, O., Hakayuwa, C.M., Obianke, A. et al. (2025) 'Diagnostic accuracy of ECG smart chest patches versus PPG smartwatches for atrial fibrillation detection: A systematic review and meta-analysis', BMC Cardiovascular Disorders, 25, 132. https://doi.org/10.1186/s12872-025-04582-2

30. Staerk, L., Sherer, J.A., Ko, D., Benjamin, E.J. and Helm, R.H. (2017) 'Atrial fibrillation: Epidemiology, pathophysiology, and clinical outcomes', Circulation Research, 120(9), pp. 1501–1517. https://doi.org/10.1161/CIRCRESAHA.117.309732

31. Šimėnas, S., Klepachevskyi, D., Medvedev, V., Treigys, P., Abramikas, Ž., Bernataviičienė, J.: Inter-patient evaluation of diffusion-based ECG denoising under standardized noise types. (2025)

32. Tripathi, A. et al. (2022) 'Computational methods for ECG denoising and AFib detection: A review', Archives of Computational Methods in Engineering. https://doi.org/10.1007/s11831-021-09659-5

33. Tsutsui, K., Biton Brimer, S., Ben-Moshe, N., Sellal, J.-M., Oster, J., Mori, H., Ikeda, Y., Arai, T., Nakano, S., Kato, R. and Behar, J.A. (2025) 'SHDB-AF: A Japanese Holter ECG database of atrial fibrillation', Scientific Data, 12, 454. https://doi.org/10.1038/s41597-025-04777-4

34. Van Gelder, I.C., Rienstra, M., Bunting, K.V. et al. (2024) '2024 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS', European Heart Journal, 45(36), pp. 3314–3414. https://doi.org/10.1093/eurheartj/ehae176

35. Vektarytė, D., Juškevičius, Z. K., & Bernatavičienė, J. (2026). An experimental study of atrial fibrillation detection using deep neural networks under noisy conditions. Lietuvos

magistrantų informatikos ir IT tyrimai, 317–325. https://www.journals.vu.lt/open-series/lt/article/view/46777

36. Wagner, G.S. (2008) Marriott's Practical Electrocardiography. 12th edn. Philadelphia, PA: Lippincott Williams & Wilkins.

37. Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F.I., Samek, W. and Schaeffter, T. (2020) 'PTB-XL, a large publicly available electrocardiography dataset', Scientific Data, 7(1), 154. https://doi.org/10.1038/s41597-020-0495-6

38. Wu, X., Yan, M., Wang, R. and Xie, L. (2025) 'MFEG-Net: Multiscale feature enhanced gating network for atrial fibrillation detection', Computer Methods and Programs in Biomedicine, 261, 108606. https://doi.org/10.1016/j.cmpb.2025.108606

39. Xia, Y., Wulan, N., Wang, K. and Zhang, H. (2018) 'Detecting atrial fibrillation by deep convolutional neural networks', Computers in Biology and Medicine, 93, pp. 84–92. https://doi.org/10.1016/j.compbiomed.2017.12.007

40. Xie, J., Stavrakis, S. and Yao, B. (2024) 'Automated identification of atrial fibrillation from single-lead ECGs using multi-branching ResNet', Frontiers in Physiology, 15, 1362185. https://doi.org/10.3389/fphys.2024.1362185

41. Zhang, D., Wang, S., Li, F., Wang, J., Sangaiah, A.K., Sheng, V.S. and Ding, X. (2019) 'An ECG signal de-noising approach based on wavelet energy and sub-band smoothing filter', Applied Sciences, 9(22), 4968. https://doi.org/10.3390/app9224968

42. Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H. and Rakovski, C. (2020) 'A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients', Scientific Data, 7(1), 48. https://doi.org/10.1038/s41597-020-0386-x

43. Zihlmann, M., Perekrestenko, M. and Tschannen, R. (2017) 'Convolutional recurrent neural networks for electrocardiogram classification', in 2017 Computing in Cardiology (CinC). https://doi.org/10.22489/CinC.2017.070-060

44. Zimetbaum, P. (2017) 'Atrial fibrillation', Annals of Internal Medicine, 166(5), ITC33–ITC48. https://doi.org/10.7326/AITC201703070

SUMMARY

Atrial fibrillation (AFib) is a major risk factor for stroke, but it may be not detected in time due to being asymptomatic. Wearable devices and automatic AFib detection algorithms enable continuous monitoring and earlier detection. Numerous deep learning models have been proposed, but direct comparison between them is difficult due to different experimental protocols used in studies. Also their robustness to noise and use of diffusion-based denoising methods remain underexplored. Therefore, we evaluated three deep learning architectures: CTRhythm, MFEGNet, and MGCNet under clean and noisy conditions and tested the effect of diffusion-based denoiser DeScoD-ECG effect on AFib classification accuracy.

These classifiers were trained on SHDB-AF dataset and cross tested on AFDB. CTRhythm, achieved the highest classification performance. Further experiments with different window lengths and loss functions identified 10-second windows and focal loss as optimal for model training, which resulted in macro F1 score 0.875 on external dataset.

Noise robustness experiments revealed severe performance degradation across all noise types. Diffusion-based denoising recovered model performance for BW noise, and restored macro F1 from 0.561 to 0.954 on SHDB-AF dataset. However, denoising was limited for MA and EM noise. These findings highlight that classification performance on clean signals does not guarantee results on noisy signals, and that noise evaluation and robust denoising steps are an important part of any AFib detection pipelines.

SUMMARY IN LITHUANIAN

Prieširdžių virpėjimas (PV) yra vienas iš pagrindinių insulto rizikos veiksnių, tačiau dažnai lieka nenustatytas dėl asimptominės eigos. Nešiojamieji įrenginiai leidžia nuolat stebėti širdies ritmą, o automatiniai PV aptikimo algoritmai gali padėti nustatyti PV laiku. Tiesiogiai palyginti egzistuojančius giliojo mokymosi algortimus yra sudentiga dėl to, kad straipsniuose yra naudojami skirtingi tyrimo protokolai. Taip pat algoritmų atsparumas triukšmui nėra pakankamai ištirtas. Todėl šiame darbe buvo įvertinti trys giliojo mokymosi modeliai: CTRhythm, MFEGNet ir MGCNet, naudojant švarius ir triukšmingus signalus. Taip pat buvo ištirtas difuzija grįsto nutriukšminimo modelio DeScoD-ECG poveikis PV klasifikavimo rezultatams.

Visi trys klasifikatoriai apmokyti naudojant SHDB-AF duomenų rinkinį ir įvertinti naudojant treniravime nenaudotą AFDB duomenų rinkinį. Iš trijų klasifikatorių geriausius rezultatus pasiekė CTRhythm. Eksperimentai su skirtingais lango ilgiais ir nuostolių funkcijomis parodė, kad 10 sekundžių langas ir židinio nuostolių funkcija pagerino klasifikavimo rezultatą ir leido pasiekti 0,875 makro F1 įvertį.

Klasifikavimo rezultatai pablogėjo esant visų tipų triukšmui. Tačiau difuzinis modelis sugebėjo atkurti signalą ir makro F1 įvertis padidėjo nuo 0,567 iki 0,954 įvertinimui naudojant nutriukšmintą SHDB-AF duomenų rinkinį. Tačiau raumenų artefaktų ir elektrodų triukšmo atveju nutriukšminimo rezultatai buvo prastesni. Šie rezultatai parodo, kad triukšmo slopinimo algoritmų naudojimas yra svarbus PV klasifikacijos uždaviniams.