# Machine Intelligence on the Edge: Interpretable Cardiac Pattern Localisation Using Reinforcement Learning

Haozhe Tian[1†], Qiyu Rao[2†], Nina Moutonnet[3], Pietro Ferraro[1] and Danilo Mandic[2]

[1]Dyson School of Design Engineering, Imperial College London, London SW7 2DB, UK.
[2]Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK.
[3]Department of Computing, Imperial College London, London SW7 2RH, UK.

[†]These authors contributed equally to this work.

## Abstract

Matched filters are widely used to localise signal patterns due to their high efficiency and interpretability. However, their effectiveness deteriorates for low signal-to-noise ratio (SNR) signals, such as those recorded on edge devices, where prominent noise patterns can closely resemble the target within the limited length of the filter. One example is the ear-electrocardiogram (ear-ECG), where the cardiac signal is attenuated and heavily corrupted by artefacts. To address this, we propose the Sequential Matched Filter (SMF), a paradigm that replaces the conventional single matched filter with a sequence of filters designed by a Reinforcement Learning agent. By formulating filter design as a sequential decision-making process, SMF adaptively design signal-specific filter sequences that remain fully interpretable by revealing key patterns driving the decision-making. The proposed SMF framework has strong potential for reliable and interpretable clinical decision support, as demonstrated by its state-of-the-art R-peak detection and physiological state classification performance on two challenging real-world ECG datasets. The proposed formulation can also be extended to a broad range of applications that require accurate pattern localisation from noise-corrupted signals.

**Keywords:** Reinforcement Learning, Pattern Localisation, Matched Filter, Wearable Sensors, ECG R-peak Detection

# 1 Introduction

Pattern localisation is a fundamental problem in signal processing, with applications spanning biomedicine, radar, and finance [1–4]. The recent development of edge signal acquisition devices has created a strong demand for more robust pattern localisation methods capable of tackling low signal-to-noise ratio (SNR) conditions. One prominent example of pattern localisation on edge is detecting (localising) R-peaks—the patterns corresponding to the electrical activity of ventricular depolarisation—in ear-electrocardiograms (ECGs), which are recorded via electrodes placed in the ear canal instead of on the chest [5]. Whilst being more convenient to set up and record over a prolonged period of time, the ear-ECG signal suffers from an extremely low SNR in comparison to the chest ECG. This is due to the attenuation of the ECG signal during propagation to the ear canal, plus the ear's proximity to non-cardiac sources of noise, including major arteries, muscles, and brain activity. Consequently, despite its central role in advanced physiological monitoring [6–10], R-peak localisation in ear-ECG remains challenging, with existing methods failing to deliver reliable performance (see Fig. 1).
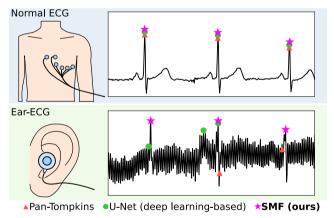


**Fig. 1** Despite the clear advantages in convenience, the critically low SNR of ear-ECG leads to poor R-peak detection performance by both the widely used Pan–Tompkins algorithm [11] and the state-of-the-art method using the U-Net [12]. In contrast, our proposed method, SMF, exhibits robust localisation performance on both normal ECG and the challenging ear-ECG.

Outside of ear-ECG, advanced pattern localisation methods have been applied to detect R-peaks in low-SNR ECG signals. These methods often rely on deep learning (DL), where neural networks are trained to minimise a proxy loss function, such as the Binary Cross-Entropy (BCE) loss, so that they predict whether each sample in an ECG segment corresponds to an R-peak [13–15]. The recent advances in neural network architectures, such as the U-Net [16], have further improved the performance of DL-based R-peak detection methods [12]. However, these approaches suffer from poor interpretability,

as it is unclear which signal patterns drive the localisation decisions. Moreover, optimising the proxy loss function does not fully align with clinically relevant performance metrics, such as the true positive, false positive, and false negative rates of R-peak detection.

An alternative, more interpretable approach is the matched filter (MF) [17, 18], which exploits the characteristic QRS pattern around R-peaks. After correlating a QRS-like template with a noisy ECG signal, high correlation values appear at the target QRS pattern locations (i.e., the R-peaks), whereas low values occur elsewhere due to morphological differences between the template and noise [19]. However, MF is inherently limited in differentiating true R-peaks from artefacts with high prominence and similar morphology. This challenge is particularly acute in ear-ECG, where the amplitudes of artefact peaks are comparable to or even exceed the target R-peaks. Moreover, existing MF templates are typically manually defined or derived from historical QRS patterns. Although recent work has explored learning the MF template [20], the resulting template remains static at deployment, making it suboptimal for ECGs with non-stationary QRS patterns, such as those seen in arrhythmia patients [21].

In this work, we propose the Sequential Matched Filter (SMF), which replaces the conventional single-MF paradigm with a sequence of signal-specific filters to achieve robust performance while retaining full interpretability. To automatically design the filter sequences, SMF performs sequence-level planning using a Reinforcement Learning (RL) agent, a data-driven approach that leverages recent advances in deep neural networks to achieve competitive performance in complex sequential decision-making tasks [22–26]. SMF offers several key advantages over existing pattern localisation methods:

- SMF distinguishes the target pattern from prominent noise patterns with similar waveforms by iteratively exploiting subtle morphological differences.
- SMF adapts to non-stationary target patterns, such as those observed in arrhythmia patients, by tailoring MF templates for each signal segment.
- SMF offers full interpretability by revealing, at each step, the key signal patterns that inform the final localisation decision.
- SMF outperforms existing DL-based methods by directly optimising localisation performance metrics without a proxy loss function.

To the best of our knowledge, SMF is the first method to automate sequentially applied MFs for pattern localisation. Its RL agent employs a lightweight neural network with around $150\,\mathrm{k}$ parameters ($\approx 0.57\,\mathrm{MB}$) that is suitable for real-time signal analysis on the edge [27], yet shows substantial performance gains over state-of-the-art approaches while preserving full interpretability. The competitive performance of SMF is empirically verified on two challenging real-world datasets: the noisy *ear-ECG* dataset (recorded using our custom-built in-ear sensors) and the pathological *arrhythmia ECG* dataset (recorded using handheld edge devices). Additionally, from the detected R-peaks, we derive Heart Rate Variability (HRV) features to classify physiological states,
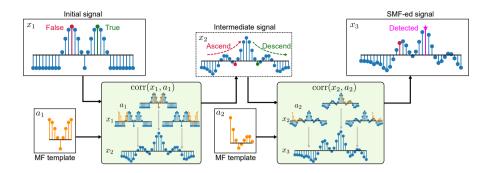
**Fig. 2** Overcoming the limitation of single-stage MF using a sequence of MFs. No single-stage MF can distinguish between the false and the true peaks in $x_1$, as the patterns surrounding them are identical within the span of the length-9 MF template. However, a sequence of only two MFs can accurately localise the true peak: the first MF, $a_1$, introduces pattern variation in $x_2$, and the second MF, $a_2$, correlates with the descending pattern that corresponds to the true peak.

where SMF exhibits robust performance, highlighting its potential for real-world cardiac health monitoring. We believe that the RL agent-driven filter design paradigm presented in this paper extends beyond the biomedical signal processing domain to a wide range of signal processing tasks that demand lightweight yet robust pattern localisation.

## 2 Preliminary

In this section, we describe the MF and present an example illustrating its inherent limitation in distinguishing between similar patterns. For a signal $x_t \in \mathbb{R}^L$, applying a MF with template $a_t \in \mathbb{R}^H$ produces the filtered signal, $x_{t+1} \in \mathbb{R}^L$, whose $n$-th sample is calculated as

$$x_{t+1}(n) = \sum_{k=0}^{H-1} a_t(k) x_t(n + k - \left\lfloor \frac{H}{2} \right\rfloor), \tag{1}$$

where $a_t(i) = 0$ when $i < 0$ or $i \geq H$, $x_t(i) = 0$ when $i < 0$ or $i \geq L$, and the operator $\lfloor i \rfloor$ is the greatest integer less than or equal to $i$. Equation (1) can be interpreted as sliding the template $a_t$ over the signal $x_t$ and measuring their correlation at these positions (see illustrations in Fig. 2, green blocks). Ideally, the correlation level, $x_{t+1}$, should peak at the index of the target pattern, allowing straightforward pattern localisation by locating the prominent maxima. However, MF performance degrades when noise contains patterns similar to the target, as illustrated in the example in Fig. 2, where distinguishing the true peak (green, right) from the false peak (red, left) in $x_1$ is infeasible for any template $a_t$ with limited length ($H = 9$ in this case), since

the patterns surrounding both peaks are identical within the span of the template. Although a template can be designed to match the transition from the true peak to the negative baseline, the resulting correlation maxima would be delayed, introducing a temporal shift in the identified peak location.

On the other hand, the true peak can be localised with a sequence of merely two strategically designed MFs. As noted previously, a single-stage MF cannot distinguish between the false and true peaks in Fig. 2 because both appear identical within the span of the filter. However, by designing a template $a_1$ that matches the valley shape between the two peaks, the resulting output $x_2$ ascends around the false peak and descends around the true peak, thereby producing distinctive pattern variations between the noise and the target. A second template, $a_2$, is then shaped to match the descending pattern of $x_2$ near the true peak, enabling accurate extraction of the true peak. Consequently, $x_3$ exhibits its largest amplitude at the time index of the true peak. As illustrated in Fig. 2, the overall output $x_3$ indeed peaks at the true peak location, confirming the successful localisation of the true peak.

# 3 Methodology

## 3.1 SMF as a Sequential Decision-Making Process

The example in Section 2 illustrates the effectiveness of strategically designing and iteratively applying sequence-level optimised MFs. To this end, we model SMF as a sequential decision-making process involving interactions between two key components: an *environment* and an *agent*. An *episode* consists of $N$ cascaded MF steps, where the output of each step becomes the input to the subsequent step, as illustrated in Fig. 3 for $N = 4$. In the first step, the environment is initialised with an ECG segment that is randomly sampled from the training set during training, or taken directly from real-time ECG data during deployment. At each step, the RL agent takes in the environment signal (state) and generates an MF template (action), which is correlated with the environment signal. The correlation output replaces the previous environment signal. In the next step, the same procedures are repeated. After the last ($N$-th) step, R-peaks are localised from the environment's stored signal by identifying local maxima that exceed a threshold of 0.5 and are separated by at least 30 samples. During training, these localised R-peaks are compared with the ground-truth R-peaks to calculate a performance metric, which is used as a reward signal that informs the learning of the RL agent. During testing or deployment, where ground-truth R-peaks are unavailable, SMF automatically detects the R-peaks without external guidance.

Formally, the SMF framework is a Markov Decision Process (MDP), because the next state, i.e., the MF correlation output, is determined only by the current signal and the MF template applied. We denote the MDP as $\mathcal{M} = (\mathcal{X}, \mathcal{S}, \mathcal{A}, r, \text{corr})$. The signal space $\mathcal{X} \subseteq \mathbb{R}^L$ contains both the original and SMF-transformed ear-ECG signals. In this work, we set the length of collected ECG segments to $L = 250$. The state space is a set $\mathcal{S} = \{s_t = (x_t, t) \mid$
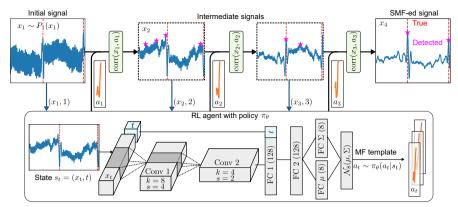
**Fig. 3** The workflow of the SMF for R-peak detection. The initial signal $x_1$ is sampled from the ECG dataset with distribution $P_1(x)$. At time step $t$, the state $s_t = (x_t, t)$ contains the signal $x_t$ and the time step $t$, based on which the RL agent generates the MF template, $a_t$, for calculating $x_{t+1}$, which is iteratively used as the state for step $t + 1$. The RL agent has policy $\pi_\theta$ in the form of a neural network that generates stochastic $a_t \sim \pi_\theta(a_t \mid s_t)$. After training, SMF templates $a_t$ are interpretable, as they reveal key signal patterns at each step.

$x_t \in \mathcal{X}, t \in \{1, \ldots, N\}\}$ that contains states $s_t$. Each $s_t \in \mathcal{S}$ is a tuple of signal $x_t$ and its corresponding time step $t$ within an episode. The action space is a set $\mathcal{A} \subseteq \mathbb{R}^H$ that contains all possible MF templates. In this work, we set $H = 8$, which is a short enough length for most edge applications. We define the reward function $r : \mathcal{S} \to \mathbb{R}$ as

$$\begin{aligned} r(s_t) &= \delta_{tN} f(\text{TP}, \text{FP}, \text{FN}) \\ &= \delta_{tN}(10\text{TP} - 5\text{FP} - 5\text{FN}), \end{aligned} \tag{2}$$

where the $\delta_{ab} = 1$ if $a = b$ and $\delta_{ab} = 0$ if $a \neq b$. For the last state $s_N$ in the episode, the SciPy function find_peaks() [28], with height parameter set to 0.5 and distance parameter set to 30, was used to identify all eligible peaks. These identified peaks are compared against the ground-truth peaks with a tolerance of 5 time steps (0.02s). The true positive (TP) is the number of correctly identified R-peaks, the false positive (FP) is the number of falsely identified peaks, and the false negative (FN) is the number of missed peaks. As shown in (2), a positive reward is given to a high TP, while penalties are given to a high FP and a high FN. The correlation function corr : $\mathcal{X} \times \mathcal{A} \to \mathcal{X}$ is defined as $x_{t+1} = \text{corr}(x_t, a_t)$, where the $n$-th sample of $x_{t+1}$ is calculated using (1). Although we deviate from the standard MDP definition by introducing an additional internal signal space $\mathcal{X}$, the expression in (1) still ensures the core Markov property, i.e., the upcoming state of SMF depends only on the current state and the current action.

We define the stochastic SMF policy as $\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which takes an state $s_t$ as input and outputs an MF template distribution $\pi(a_t \mid s_t)$. The stochastic $\pi$ explores various MF template sequences during training, thus

avoiding getting stuck with suboptimal solutions. The optimisation objective is to maximise the expected cumulative reward in an episode, that is

$$\pi^{\star} = \operatorname*{argmax}_{\pi} \mathbb{E}_{s_1,a_1,\cdots} \left[ \sum_{t=1}^{N} r(s_t) \right], \tag{3}$$

where $s_t = (x_t, t)$, $x_1$ is sampled from the ECG dataset with $x_1 \sim P_1(x_1)$, $x_{t+1} = \operatorname{corr}(x_t, a_t)$, and $a_t \sim \pi(a_t \mid s_t)$. For the value function, $V^{\pi}$, and the state-action value function, $Q^{\pi}$, which will be used later for training the policy $\pi$, we follow the standard definitions from [29], given by:

$$V^{\pi}(s_t) = \mathbb{E}_{a_t,s_{t+1},a_{t+1},\cdots} \left[ \sum_{i=t}^{N} r(s_i) \right],$$

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{s_{t+1},a_{t+1},\cdots} \left[ \sum_{i=t}^{N} r(s_i) \right]. \tag{4}$$

## 3.2 Optimising SMF with RL

To address the continuous state space $\mathcal{S}$ and action space $\mathcal{A}$ in SMF, this study employs deep RL, where the policy $\pi$ is approximated using a neural network $\pi_\theta$ parametrized by $\theta$ to allow generalization to unobserved states and actions in $\mathcal{S}$ and $\mathcal{A}$. The policy $\pi_\theta$ takes in a state tuple $s_t = (x_t, t)$ of size $250 + 1$, and outputs an MF template of length 8, using the neural network illustrated in the lower panel of Fig. 3. The input state $s_t = (x_t, t)$ comprises a signal $x_t$ of length $L = 250$ and a scalar time index $t$. The signal $x_t$ is processed by two 1-dimensional Convolutional Neural Network (CNN) layers (Conv): the first with a kernel size of $k = 8$ and stride $s = 4$, and the second with a kernel size of $k = 4$ and stride $s = 2$. The convolution output is then transformed into a feature vector of size 128 using a fully connected layer (FC). This feature vector is concatenated with the scalar, $t$, resulting in a combined representation of size 129. This concatenated representation is further transformed by an FC layer into a vector of size 128, which is subsequently mapped through two separate FC layers to generate the mean vector $\mu$ (size 8) and the diagonal covariance matrix $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_8)$. Together, $\mu$ and $\boldsymbol{\Sigma}$ define a multivariate Gaussian distribution $\mathcal{N}_8(\mu, \boldsymbol{\Sigma})$, from which the MF template, $a_t$, is sampled using the reparametrization trick [30].

The general procedure for training $\pi_\theta$ is summarised in Algorithm 1. Most existing deep RL algorithms can be used to solve the optimisation problem in (3), among which the two most prominent categories are: i) policy gradient methods, which directly optimise $\pi_\theta$ using the objective in (3); and ii) actor–critic methods, which optimise $\pi_\theta$ through estimated state–action values, enabling learning from past experience and thereby improving sample

---

**Algorithm 1** Training SMF

---

1: **Initialization:** Train set $\mathcal{X}_{\text{train}}$ containing ECG segments and ground-truth R-peak positions, episode length $N$
2: **Output:** Trained SMF policy $\pi_\theta$
3: **repeat**
4:      Randomly sample $(x_1, \{\text{peaks}\}) \in \mathcal{X}_{\text{train}}$
5:      $t \leftarrow 1$
6:      **while** $t \leq N$ **do**
7:          Get MF template: $a_t \sim \pi_\theta(a_t \mid s_t = (x_t, t))$
8:          $x_{t+1}(n) = \sum_{k=0}^{H-1} a_t(k) x_t(n + k - \lfloor \frac{H}{2} \rfloor)$
9:          **if** $t == N$ **then**
10:            Find local maximums $\{\text{preds}\}$.
11:            Compute TP, FP, and FN by comparing $\{\text{preds}\}$ and $\{\text{peaks}\}$
12:            $r(x_t) = 10\text{TP} - 5\text{FP} - 5\text{FN}$
13:          **else**
14:            $r(x_t) = 0$
15:          **end if**
16:          $s_{t+1} = (x_{t+1}, t + 1)$
17:          Use $(s_t, a_t, s_{t+1}, r(s_t))$ to update $\pi_\theta$ (e.g., using PPO or SAC).
18:          $t = t + 1$
19:      **end while**
20: **until** *convergence* is *true*

---

efficiency. This work proposes two SMF implementations based on two state-of-the-art RL algorithms, each representing one of the two abovementioned categories.

### 3.2.1 SMF-PPO

Proximal Policy Optimisation (PPO) is a prominent policy gradient method that clips the objective to prevent excessive policy updates, thereby stabilising training [31]. The SMF-PPO first collects a batch of SMF episodes using $\pi_{\theta_{\text{old}}}$, then performs update by maximising the following objective, as

$$\max_\theta \hat{\mathbb{E}}_{s_t, a_t} \left[ L(s_t, a_t, \theta_{\text{old}}, \theta) \right], \text{ where}$$

$$L(s_t, a_t, \theta_{\text{old}}, \theta) = \begin{cases} \min \left( \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1 + \epsilon \right) \hat{A}_t, \ \hat{A}_t > 0 \\ \max \left( \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1 - \epsilon \right) \hat{A}_t, \ \hat{A}_t \leq 0 \end{cases},$$

$$\hat{A}_t = r(s_N) - V_\psi(s_t). \tag{5}$$

At state $s_t$, $\hat{A}_t$ captures how much better (or worse) each chosen MF template, $a_t$, performed compared to the expected average. The performance of $a_t$ is evaluated as the final peak extraction performance $r(s_N)$ in the episode that

includes $s_t$. The expected average is estimated using the value function, $V^\pi$, defined in (4). A positive $\hat{A}_t$ indicates that MF template $a_t$ outperforms the average of $\pi_{\theta_{\text{old}}}$. Therefore, to maximize $L(s_t, a_t, \theta_{\text{old}}, \theta)$, it is favourable to increase the probability of using $a_t$, i.e., increase $\pi_\theta(a_t \mid s_t)$. However, the $\min(\cdot)$ ensures that the updated $\pi_\theta$ is not too far from the old $\pi_{\theta_{\text{old}}}$, i.e., $\pi_\theta(a_t \mid s_t) \leq (1+\epsilon)\pi_{\theta_{\text{old}}}(a_t \mid s_t)$. The same logic applies to the case where $\hat{A}_t$ is negative. The clipping in SMF-PPO avoids drastic updates on MF templates, which could result in catastrophic performance degradations.

In practice, $V^\pi$ is approximated by a neural network $V_\psi$, which shares a similar architecture with the policy network $\pi_\theta$ but outputs a scalar value. The parameters $\psi$ are updated by minimising the mean squared error between $V_\psi$ and its Bellman estimation

$$\min_\psi \hat{\mathbb{E}}_{s_t, s_{t+1}} \left[ (V_\psi(s_t) - (r(s_t) + V_\psi(s_{t+1}))^2 \right].$$

In SMF-PPO, the estimation of the advantage $\hat{A}_t$ in (5) is further stabilised using the Generalised Advantage Estimator (GAE) [32].

### 3.2.2 SMF-SAC

Soft Actor-Critic (SAC) is a prominent actor-critic method that uses an entropy regularisation term in its objective to encourage diverse and robust policies [30]. The SMF-SAC updates its policy at every SMF step using a sampled batch of historical MF steps, each denoted by $(s_t, a_t, s_{t+1}, r(s_t))$. SMF-SAC parametrises the state-action value function $Q^\pi$ with a neural network, $Q_\phi$, with parameter $\phi$, to allow generalisation to unobserved state-action pairs. The network structure of $Q_\phi$ is similar to the policy network $\pi_\theta$ but outputs a scalar value. The $\phi$ is updated by minimizing the mean square error between $Q_\phi(s_t, a_t)$ and its Bellman estimator $r(s_t) + Q_\phi(s_{t+1}, a_{t+1})$

$$\min_\phi \hat{\mathbb{E}}_{s_t, a_t, s_{t+1}} \left[ (Q_\phi(s, a) - y)^2 \right], \text{ where}$$
$$y = r(s_t) + Q_\phi(s_{t+1}, a_{t+1}) - \alpha \log \pi(a_{t+1} \mid s_{t+1}),$$
$$a_{t+1} \sim \pi(a_{t+1} \mid s_{t+1}),$$

where the entropy regularisation term $\log \pi(a_{t+1} \mid s_{t+1})$ introduced by [30] encourages the exploration of diverse MF templates and avoids sticking to sub-optimal templates. The hyperparameter $\alpha$ adjusts the intensity of regularisation.

To update the actor $\pi_\theta$, observe that the estimated state-action value $Q_\phi$ indicates the expected future cumulative rewards of taking $a_t \sim \pi_\theta(a_t \mid s_t)$ at $s_t$, which the updated policy aims to maximise. Therefore, the $\pi_\theta$ is directly updated by maximizing the expected value of $Q_\phi$, as

$$\max_\theta \hat{\mathbb{E}}_{s_t} \left[ Q_\phi(s_t, a_t) - \alpha \log \pi(a_{t+1} \mid s_{t+1}) \right], a_t \sim \pi_\theta(a_t \mid s_t).$$
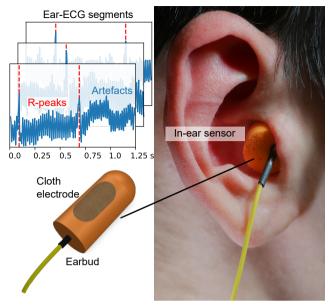
**Fig. 4** The ear-ECG acquisition setup. The ear-ECG signal was recorded using our custom-built in-ear sensors, with the right ear used for recording and the left ear serving as the reference. The in-ear sensors consist of an earbud and a soft cloth electrode. The ear-ECG signals were recorded at a sampling rate of 200 Hz and were split into segments with 250 samples (1.25s).

# 4 Hardware and Data Acquisition

The ear-ECG dataset in this study was recorded from 7 healthy subjects (5 males and 2 females, aged 20-30) under the ethics protocol JRCO 20IC6414. Our custom-built in-ear sensors were placed in both ear canals, recording signals from the right ear while using the left ear as the reference. Figure 4 depicts the in-ear sensor based on the design described in [33], which consists of an earbud and a soft cloth electrode. The earbud was made of viscoelastic foam to alleviate artefacts arising from mechanical deformations of the ear canal. The cloth electrode was made of stretchable, low-impedance fabric, ensuring reliable skin contact and enhanced comfort during prolonged wear. To further reduce impedance caused by poor skin contact, conductive gel was applied before placing the sensor in the ear canal. The signal was recorded at a sampling rate of 200 Hz and was subsequently divided into non-overlapping 250-sample segments (1.25s). The left panel of Fig. 4 presents some sample ear-ECG segments recorded using our setup. The identification of true R-peaks is extremely challenging in these ear-ECG segments due to the numerous false artefact peaks around R-peaks. The prominent false peaks with similar or even greater amplitude compared to true R-peaks pose significant challenges for accurate R-peak detection.

# 5 Experiment Setup

## 5.1 Datasets

We validated the SMF on two real-world ECG datasets:

1. The *ear-ECG* dataset contains 720 ear-ECG segments collected using the setup described in Section 4. The Hearable setup led to low signal amplitude and prominent artefact peaks, making accurate R-peak detection particularly challenging (see example in Fig. 6a).
2. The *arrhythmia ECG* dataset contains 771 single-lead ECG segments derived from subjects with atrial fibrillation, the most common cardiac arrhythmia, in the 2017 Computing in Cardiology Challenge [34]. This dataset is challenging because arrhythmia causes non-stationary QRS patterns and irregular R-R intervals (the distance between successive R-peaks), as illustrated in Fig. 6b. Additionally, the arrhythmia ECG dataset, recorded using the handheld AliveCor device, is relatively noisy and susceptible to electrode misplacement, which can invert the signal in some segments (see the left panel of Fig. 6c).

For both the ear-ECG dataset and the arrhythmia ECG dataset, each ECG segment contained 250 samples (1.25s). We used 70% of the ECG segments for training and the remaining 30% for testing. The train-test split remained consistent across all experiments. The test set was strictly reserved for evaluation, ensuring that neither the SMF nor the baseline methods could access it during training. To train the SMF, we designed two RL environments in the widely used OpenAI Gym style [35], corresponding to the ear-ECG and arrhythmia ECG datasets. During training, the RL environments were reset every $N$ steps with a randomly selected ECG segment from the training set. During testing, the N-step SMF was applied to all ECG segments in the test set for calculating the average performance metrics.

## 5.2 Implementations of SMF and Baselines

To train the proposed SMF method, $10^5$ SMF steps were observed. For SMF-PPO, the policy $\pi_\theta$ was updated after every 500 consecutive transitions. For each transition, the cumulative reward used for advantage estimation was the final R-peak detection performance of the same episode. The 500 collected single-step transitions were then randomly split into four mini-batches of 125 transitions. The policy update was performed over four epochs, with each epoch iterating through all mini-batches. The learning rate was $10^{-4}$ and the clipping ratio $\epsilon$ was set to 0.2. Gradient clipping was applied to avoid gradients larger than 0.5. For SMF-SAC, the policy $\pi_\theta$ was updated every 2 steps using a batch of 512 historical single-step transitions stored in a Replay Buffer [36]. A Polyak weight averaging with a smoothing factor 0.005 was used to stabilise the update of $Q$ networks [37]. The learning rate was set to $10^{-4}$ for both policy updates and $Q$ network updates. The entropy regularisation term $\alpha$ was

set to 0.2. The training setup and hyperparameters were kept the same for all experiments in this section.

The neural network architectures used by SMF are lightweight and well-suited for edge deployment. For SMF-PPO, sharing parameters between the value function and the policy results in an RL agent with approximately $156\,\mathrm{k}$ parameters ($\approx0.60\,\mathrm{MB}$). For SMF-SAC, both the $Q$-network and the policy network contain around $140\,\mathrm{k}$ parameters ($\approx0.54\,\mathrm{MB}$). As demonstrated in our previous work [27], these RL agents can be readily deployed on edge devices to achieve real-time pattern localisation, requiring only milliseconds to process 60-second ECG recordings on an Android smartphone.

To evaluate the performance and robustness of SMF, we empirically compared it against the following baselines.

1. The *Pan-Tompkins* algorithm [11] is arguably the most widely used R-peak detection method that combines a series of complex signal processing operations and a decision rule that decides the validity of each potential R-peak by comparing the current R-R interval to the average of historical R-R intervals.
2. The *Bidirectional RNN (Bi-RNN)* [38] is a popular sequential neural network architecture that uses both forward and backward recurrent layers to capture context from past and future signals.
3. The *U-Net*[12, 16] is a CNN-based neural network architecture that [12] reports achieving state-of-the-art performance in ECG R-peak detection.
4. The *MF-PPO* and *MF-SAC* are ablated versions of the proposed SMF algorithm, restricted to episode lengths of 1. We include them as baselines to represent the performance of single-stage MFs when optimised in a data-driven manner. Although non-sequential, these baselines generate MF templates that directly optimise R-peak detection performance by using the RL reward function in (2) as their objective.

The Bi-RNN baseline consisted of two Bi-RNN layers with hidden sizes of 64, which transformed the signal into a feature vector of length 250 with 128 channels, followed by a linear layer that mapped this feature vector to a length 250 prediction vector. The U-Net baseline followed [12]. As discussed in Section 1, it is infeasible to directly optimise TP, FP, and FN with DL-based methods. Therefore, the objectives of the DL methods were to minimise the binary cross-entropy (BCE) loss between the network prediction vector and a binary vector of length 250, where ones correspond to R-peak positions and zeros to non-peak positions. Both Bi-RNN and U-Net were trained over 1000 epochs with a batch size of 100 and a learning rate of 0.005.

# 6 Results and Analysis

## 6.1 Comparison of R-peak Detection Performance.

We first evaluated the best R-peak detection performance. The SMF-PPO and SMF-SAC had episode lengths of 3, i.e., they automatically applied 3 iterative

**Table 1** R-peak detection performance

| Method | | Precision | Recall | F-1 score |
|--------|--|-----------|--------|-----------|
| Ear-ECG | | | | |
| Pan-Tompkins | [11] | 0.5520 | 0.5243 | 0.5378 |
| Bi-RNN | [38] | 0.7670 | 0.8778 | 0.8187 |
| U-Net | [12] | 0.9029 | 0.9490 | 0.9254 |
| MF-PPO | | 0.9900 | 0.9612 | 0.9754 |
| MF-SAC | | 0.9700 | 0.9417 | 0.9557 |
| SMF-PPO | | 0.9902 | 0.9806 | 0.9854 |
| SMF-SAC | | **1.0000** | **0.9826** | **0.9902** |
| Arrhythmia ECG | | | | |
| Pan-Tompkins | [11] | 0.6676 | 0.4962 | 0.5693 |
| Bi-RNN | [38] | 0.9567 | 0.8156 | 0.8806 |
| U-Net | [12] | 0.9338 | 0.8843 | 0.9084 |
| MF-PPO | | 0.9073 | 0.9211 | 0.9141 |
| MF-SAC | | 0.8978 | 0.9160 | 0.9068 |
| SMF-PPO | | 0.9446 | 0.9542 | 0.9494 |
| SMF-SAC | | **0.9543** | **0.9567** | **0.9555** |

MFs to optimise R-peak detection performance in the last step. The rationale for selecting an episode length of 3 is provided in Section 6.3.2. In contrast, MF-PPO and MF-SAC employed the same RL training framework but with an episode length of 1, corresponding to a single-stage MF. These were included as baselines to represent data-driven optimisation of single-stage MFs. The results are shown in Table 1, where the performance metrics are given by

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{F-1} = \frac{\text{TP}}{\text{TP} + 0.5 \times (\text{FP} + \text{FN})} \tag{6}$$

Table 1 shows that SMF-SAC consistently achieved the highest precision, recall, and F-1 scores on both datasets. The Pan-Tompkins method only achieved an F-1 score of approximately 0.55, caused by its reliance on historical R-R intervals, which are misleading in arrhythmia ECG signals with irregular R-R intervals. The DL-based methods, Bi-RNN and U-Net, underperformed SMFs. Notably, their performance fell below that of the non-sequential MF-PPO and MF-SAC, highlighting the advantage of the proposed RL paradigm, which directly optimises R-peak detection, over the existing DL paradigm, which relies on minimising proxy loss functions. We also observed that MF-PPO and MF-SAC with episode lengths of 1 underperformed SMF-PPO and SMF-SAC with episode lengths of 3, providing empirical evidence that the sequential application of MFs can overcome the inherent limitations in single-stage MF.

To evaluate the statistical significance of SMF's performance gains, Fig. 5 presents the average test set F-1 scores of the proposed SMF methods and the baseline methods, where the means and standard deviations were obtained by conducting five independent runs for each method using different random seeds. To show that the advantage of SMF-PPO and SMF-SAC was
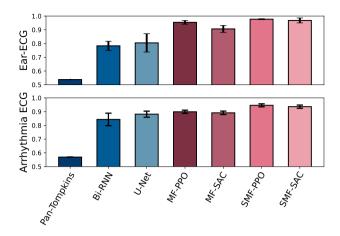
**Fig. 5** Average F-1 scores of R-peak detection, with the main bars as means and the error bars as standard deviations (where applicable).

**Table 2** The p-values in t-tests comparing SMF with the baselines.

|  | Bi-RNN | U-Net | MF-PPO | MF-SAC |
|---|---|---|---|---|
| Ear-ECG | | | | |
| SMF-PPO | 0.000 | 0.001 | 0.011 | 0.001 |
| SMF-SAC | 0.000 | 0.002 | 0.254 | 0.004 |
| Arrhythmia ECG | | | | |
| SMF-PPO | 0.002 | 0.001 | 0.000 | 0.000 |
| SMF-SAC | 0.004 | 0.002 | 0.002 | 0.001 |

statistically significant compared to the baselines, we performed t-tests and report the p-values in Table 2. Observe that SMF consistently achieved the highest average F-1 scores, with low standard deviations compared to the DL-based methods. Note that although the performance difference between SMF-SAC and MF-PPO in the ear-ECG dataset was not statistically significant ($p > 0.05$), for both SAC and PPO, their sequential version significantly outperformed their corresponding non-sequential version.

Our previous work has demonstrated that SMF neural networks can run in real time on smartphones [27]. In Table 3, we compare the average time required by SMF and the baseline methods to process a 1.25 s ECG segment. All methods were averaged over 10 runs across all test segments from the arrhythmia ECG dataset. All experiments were conducted on an Ubuntu 22.04 system with an Intel Core i7-13850HX CPU and an Nvidia RTX 3500 Ada GPU. Being faster than the widely used, real-time Pan-Tompkins method, SMF met the requirement for real-time R-peak detection. Although SMF was slightly slower than U-Net due to its sequential nature, it outperformed the Bi-RNN baseline, as its CNN-based architecture enables greater parallelisation than recurrent models.

**Table 3** Average Processing Time of 1.25 s ECG Segment ($10^{-3}$ s).

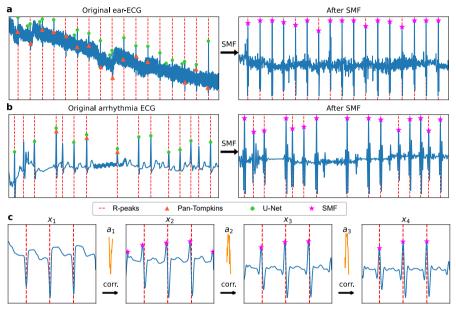| Pan-Tompkins | Bi-RNN | U-Net | SMF-PPO | SMF-SAC |
|---|---|---|---|---|
| 4.423 | 2.421 | 0.987 | 1.403 | 1.372 |



**Fig. 6** Comparison of the proposed SMF method (SMF-SAC) with the widely used Pan–Tompkins algorithm and the state-of-the-art U-Net on example ECG segments: (**a**) a noisy ear-ECG section containing numerous false peaks caused by non-cardiac artefacts; (**b**) an arrhythmia ECG section with varying R–R intervals, which challenge rule-based approaches such as Pan–Tompkins. In (**c**), we show SMF procedures applied to an arrhythmia ECG segment with inverted R-peaks due to improper recording setup. SMF automatically corrects the inversion by applying signal-aware templates, which remain interpretable (e.g., $a_1$ corresponds to the inverted R-peak morphology).

## 6.2 Visualizing the Robustness

This work focuses on R-peak detection in the Hearable setting, where a robust ECG R-peak detection method must handle prominent false artefact peaks and ECGs caused by varying patient physiology and recording devices. Figure 6a shows the performance of SMF (SMF-SAC) for a noisy ear-ECG section, where false peaks had even greater prominence than the true R-peaks. Nevertheless, SMF localises all true R-peaks. The raw ear-ECG signal also exhibited a baseline drift (see the descending trend in Fig. 6a), which was effectively removed by SMF. Figure 6b shows the performance of SMF on an arrhythmia ECG section, with a large variance in R-R intervals. The rule-based Pan-Tompkins method identified only 3 out of the 15 R-peaks, as its decision rule relies on historical R-R intervals. In contrast, SMF iteratively refined the signal to high quality, eliminating the need for decision rules based on historical R-R intervals. Figure 6c shows the complete workflow of SMF on an arrhythmia ECG

segment with inverted R-peaks caused by the misplacement of the recording ECG leads. Despite the inversions, SMF successfully localised the R-peaks by generating a series of MF templates with large negative deflections that aligned with the inverted R-peaks. The results also highlight the interpretability of SMF. Compared to the MF templates for healthy patients in Fig. 3, which exhibited more prominent positive deflections, the templates for arrhythmia patients in Fig. 6c featured more prominent negative deflections. These signal patterns encoded in SMF templates could be useful indicators for diagnosing cardiovascular diseases [39].

## 6.3 Sensitivity Analysis

We now provide more insights into the proposed SMF method by varying its episode length, reward function, and MF template length. Figure 7 plots the test set R-peak detection F-1, evaluated at different training steps. For each design choice, we trained SMF for five independent runs with different random seeds, and plotted the mean test set F-1 score as solid lines and standard deviations as shaded areas.

### 6.3.1 Impact of Sequential Applications of MFs

A key insight of this work is that iteratively applying MFs can overcome the inherent limitation of single-stage MFs. To verify this, we trained SMFs with varying episode lengths, i.e., the number of iterative MFs applied, and evaluated their R-peak detection performance (Fig. 7a). The non-sequential MF-PPO and MF-SAC with episode length 1 achieved average F-1 scores of 0.95 and 0.91, respectively. As the episode length increased, SMFs achieved higher F-1 scores on the test set, with the optimal episode length being 3, where the average F-1 scores were 0.98 and 0.97 for SMF-PPO and SMF-SAC, respectively. This demonstrates that the iterative application of MFs can overcome the limitation of non-sequential MFs. When the episode length was 4, the improvement was marginal for SMF-SAC and even negative for SMF-PPO. We also observed that, as episode length increased, SMFs required more training steps to converge. These were expected, as the sequential decision-making nature of the SMF problem means that the number of samples needed to derive an effective policy grows exponentially with the increase in episode length. Therefore, it became increasingly difficult to learn a stable policy as episode lengths increased.

### 6.3.2 Effect of Template Length

We explored how MF template length affected SMF performance. We experimented with template lengths of 4, 8, 12, and 16 samples in the arrhythmia ECG dataset (Fig. 7b). The results show that a template length of 4 yielded the worst performance for both SMF-PPO and SMF-SAC, as such a short template failed to capture distinctive R-peak patterns. Since the QRS pattern around R-peaks lasts about 0.08 seconds (16 samples) in adults, longer MF templates
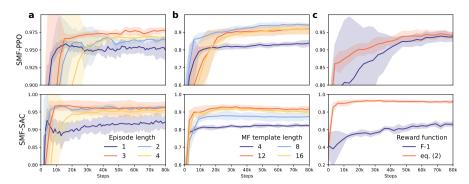
**Fig. 7** Sensitivity of the SMF method to parameter changes. The sub-figures show the test set F-1 score, evaluated at different training stages in the ear-ECG environment. The solid lines denote the means, while the shaded areas denote the standard deviations. (**a**) SMFs with different episode lengths, i.e., the number of MF iterations applied, in the ear-ECG dataset. (**b**) SMFs with different template lengths in the arrhythmia ECG dataset. (**c**) SMFs with different reward functions in the arrhythmia ECG dataset.

improve correlation with true peaks while reducing artefact correlation. However, increasing template length also exponentially expands the action space, making the sequential decision-making problem more challenging. This effect was most evident when the template length was 16, where SMF-PPO exhibited unstable training and lower converged F-1 than SMFs with shorter templates. The optimal template length was 8 for SMF-PPO and 12 for SMF-SAC. The SMF-SAC performed better with longer templates due to its higher sample efficiency, as it leverages all past transitions, whereas SMF-PPO learns only from transitions generated by the current policy. Additionally, SMF-SAC's entropy regularisation promotes exploration in a broader state-action space, resulting in more stable training and enhanced performance.

### 6.3.3 Influence of Reward Function Design

We examined the impact of the reward function design on SMF performance (Fig. 7c). For all other experiments in this section, SMF utilised the reward function in (2). To quantify the effect of reward function design, we compared this reward design with another straightforward scheme, where the F-1 score in (6) was used as the reward. For SMF-PPO, although training with the F-1 score as the reward was slower and less stable, the final converged F-1 was comparable to that obtained using the reward from (2). For SMF-SAC, using the F-1 score as the reward significantly hindered performance. As the lower panel of Fig. 7c shows, SMF-SAC with the F-1 score as the reward function achieved a test set F-1 score of only around 0.6, which was lower (by $> 0.35$) than that achieved with the reward from (2). While using the F-1 score as the reward reflects the general objective of reducing FP and FN, it does not differentiate the impact of each type of error straightforwardly. In contrast, the reward function $10\text{TP} - 5\text{FP} - 5\text{FN}$ linearly assigns different penalties to
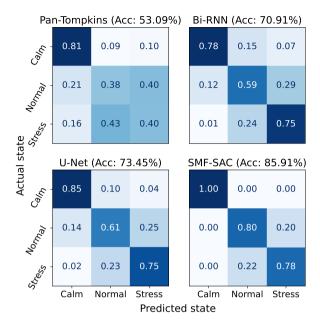
**Fig. 8** Physiological state classification based on R-peaks extracted in ear-ECG recordings. For each method, the extracted R-peaks were used to compute features for training random forest classifiers. The average accuracy and normalised confusion matrices were obtained over 100 MCCVs using a train-test split of 70:30%.

FP and FN, enabling the agent to adjust for each type of error and facilitating easier learning in the environment.

## 6.4 Physiological State Classification

As mentioned in Section 1, rich physiological information can be extracted from R-peaks in ear-ECG. To validate this, we performed a physiological state classification task using the ear-ECG signal. The classification setup was similar to that in [40], where three states were considered: calm, during which the subject performed controlled deep breathing; normal, where the subject remained seated and still; and stress, during which the subject solved mental exercises. Each state was recorded for 300 seconds and split into 25-second sections, resulting in a total of 36 sections (12 sections per state).

For both SMF and baseline methods, R-peaks were first localised, followed by the extraction of five widely used Heart Rate Variability (HRV) features: RMSSD, SDNN, HR, LF, HF, and LF/HF (for details on calculating these features, please refer to the summary in [41]). These features were then used to train random forest classifiers. To evaluate classification performance, we computed the average accuracy and normalised confusion matrices over 100 Monte Carlo Cross-Validation (MCCV) runs using a 70:30% train-test split

(Fig. 8). The results showed that features derived from SMF-SAC achieved significantly higher classification accuracy compared to baseline methods. For the calm state, SMF-extracted R-peaks led to perfect classifications, a result that the DL-based methods could not achieve. This highlights the strong potential of SMF for physiological state monitoring on the edge.

# 7 Conclusion

Beyond the improved convenience in setup and suitability for prolonged recordings, the rise of edge signal acquisition devices has also created a strong demand for robust, explainable target pattern localisation methods that support trustworthy decision-making. A prominent example is the ear-ECG signals with critically low Signal-to-Noise Ratio, where the reliable localisation of R-peaks can greatly enhance cardiac monitoring and diagnosis. This work addresses this challenge by introducing the Sequential Matched Filter (SMF), which leverages a Reinforcement Learning (RL) agent to design signal-specific filter sequences for robust and interpretable pattern localisation. The RL agent of SMF employs lightweight neural network architectures that are suitable for edge deployment. When evaluated on two challenging real-world ECG datasets, SMF achieves state-of-the-art R-peak detection performance. At the same time, it remains fully interpretable by revealing key signal patterns (e.g., the QRS patterns in ECG) at each step, thereby supporting trustworthy clinical decision-making and enabling the identification of cardiac abnormalities or sensor misplacement. Moreover, we empirically demonstrate that SMF's improved localisation performance directly enables reliable physiological state classification on the edge. An intriguing conclusion is that SMF provides a robust and interpretable digital filter design framework applicable to edge signal processing tasks beyond the biomedical domain.

# References

[1] Fu, T.-c.: A review on time series data mining. Engineering Applications of Artificial Intelligence **24**(1), 164–181 (2011)

[2] Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics **13**(6), 395–405 (2012)

[3] Li, K., Ma, Z., Robinson, D., Ma, J.: Identification of typical building daily electricity usage profiles using gaussian mixture model-based clustering and hierarchical clustering. Applied energy **231**, 331–342 (2018)

[4] Chen, Y., Lin, R., Cheng, Y., Li, J.: Joint design of periodic binary probing sequences and receive filters for pmcw radar. IEEE Transactions on Signal Processing **70**, 5996–6010 (2022)

20

[5] Xu, Y., Uppal, A., Lee, M.S., Mahato, K., Wuerstle, B.L., Lin, M., Djassemi, O., Chen, T., Lin, R., Paul, A., et al.: Earable multimodal sensing and stimulation: A prospective towards unobtrusive closed-loop biofeedback. IEEE Reviews in Biomedical Engineering (2024)

[6] Thayer, J.F., Åhs, F., Fredrikson, M., Sollers III, J.J., Wager, T.D.: A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. Neuroscience & Biobehavioral Reviews **36**(2), 747–756 (2012)

[7] Kleiger, R.E., Stein, P.K., Bigger Jr, J.T.: Heart rate variability: Measurement and clinical utility. Annals of Noninvasive Electrocardiology **10**(1), 88–101 (2005)

[8] Cole, C.R., Blackstone, E.H., Pashkow, F.J., Snader, C.E., Lauer, M.S.: Heart-rate recovery immediately after exercise as a predictor of mortality. New England Journal of Medicine **341**(18), 1351–1357 (1999)

[9] Wang, K.-K., Yang, G.-P., Yang, L., Huang, Y.-W., Yin, Y.-L.: Ecg biometrics via enhanced correlation and semantic-rich embedding. Machine Intelligence Research **20**(5), 697–706 (2023)

[10] Gendler, J., Zhou, A., Krishnamurthi, N., Rand, C., Weese-Mayer, D.: Heart rate variability as a biomarker of progressive cardiac autonomic nervous system dysregulation in congenital central hypoventilation syndrome (CCHS). American Journal of Respiratory and Critical Care Medicine **211**(Abstracts), 1273–1273 (2025)

[11] Pan, J., Tompkins, W.J.: A real-time QRS detection algorithm. IEEE Transactions on Biomedical Engineering **3**, 230–236 (1985)

[12] Zahid, M.U., Kiranyaz, S., Ince, T., Devecioglu, O.C., Chowdhury, M.E., Khandakar, A., Tahir, A., Gabbouj, M.: Robust R-peak detection in low-quality holter ECGs using 1D convolutional neural network. IEEE Transactions on Biomedical Engineering **69**(1), 119–128 (2021)

[13] Laitala, J., Jiang, M., Syrjälä, E., Naeini, E.K., Airola, A., Rahmani, A.M., Dutt, N.D., Liljeberg, P.: Robust ECG r-peak detection using lstm. Proceedings of the 35th Annual ACM Symposium on Applied Computing (ACM SAC), 1104–1111 (2020)

[14] Zhou, P., Schwerin, B., Lauder, B., So, S.: Deep learning for real-time ECG R-peak prediction. Proceedings of the 14th International Conference on Signal Processing and Communication Systems (ICSPCS), 1–7 (2020). IEEE

[15] Peng, X., Zhu, H., Zhou, X., Pan, C., Ke, Z.: ECG signals segmentation

using deep spatiotemporal feature fusion u-net for qrs complexes and R-peak detection. IEEE Transactions on Instrumentation and Measurement **72**, 1–12 (2023)

[16] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th Medical image computing and computer-assisted intervention (MICCAI), part III 18, 234–241 (2015)

[17] Hamilton, P., Tompkins, W.: Adaptive matched filtering for qrs detection. Proceedings of the 10th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 147–148 (1988). IEEE

[18] Xue, Q., Hu, Y.H., Tompkins, W.J.: Neural-network-based adaptive matched filtering for qrs detection. IEEE Transactions on Biomedical Engineering **39**(4), 317–329 (1992)

[19] Chanwimalueang, T., von Rosenberg, W., Mandic, D.P.: Enabling R-peak detection in wearable ECG: Combining matched filtering and Hilbert transform. Proceedings of the IEEE International Conference on Digital Signal Processing (DSP), 134–138 (2015)

[20] Davies, H.J., Hammour, G., Zylinski, M., Nassibi, A., Stanković, L., Mandic, D.P.: The deep-match framework: R-peak detection in ear-ECG. IEEE Transactions on Biomedical Engineering **71**(7), 2014–2021 (2024)

[21] Clifford, G.D., Azuaje, F., McSharry, P., *et al.*: Advanced Methods and Tools for ECG Data Analysis vol. 10. Artech House Boston, ??? (2006)

[22] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., *et al.*: Mastering the game of go without human knowledge. Nature **550**(7676), 354–359 (2017)

[23] Ying, M.-S., Feng, Y., Ying, S.-G.: Optimal policies for quantum markov decision processes. International Journal of Automation and Computing **18**(3), 410–421 (2021)

[24] Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D., *et al.*: Magnetic control of tokamak plasmas through deep reinforcement learning. Nature **602**(7897), 414–419 (2022)

[25] Aung, H.W., Li, J.J., An, Y., Su, S.W.: A real-time framework for EEG signal decoding with graph neural networks and reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems (2025)

[26] Chen, Y., Xiao, J.: Target search and navigation in heterogeneous robot systems with deep reinforcement learning. Machine Intelligence Research **22**(1), 79–90 (2025)

[27] Zylinski, M., Davies, H.J., Rao, Q., Mandic, D.P.: Hearables: Deep matched filter for online R-peak detection from in-ear ECG in mobile application. The Library **7**, 11 (2023)

[28] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2

[29] Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P.: Trust region policy optimization. Proceedings of the 32nd International Conference on Machine Learning, 1889–1897 (2015)

[30] Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. Proceedings of the 37th International Conference on Machine Learning, 1861–1870 (2018)

[31] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

[32] Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P.: High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438 (2015)

[33] Goverdovsky, V., Von Rosenberg, W., Nakamura, T., Looney, D., Sharp, D.J., Papavassiliou, C., Morrell, M.J., Mandic, D.P.: Hearables: Multimodal physiological in-ear sensing. Scientific Reports **7**(1), 6948 (2017)

[34] Clifford, G.D., Liu, C., Moody, B., Li-wei, H.L., Silva, I., Li, Q., Johnson, A., Mark, R.G.: AF classification from a short single lead ECG recording: The Physionet/computing in cardiology challenge 2017. Proceedings of the 2017 Computing in Cardiology (CinC), 1–4 (2017)

[35] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: Openai gym. arXiv preprint arXiv:1606.01540 (2016)

[36] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013)

[37] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015). https://doi.org/10.1038/nature14236

[38] Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997)

[39] Libby, P., Theroux, P.: Pathophysiology of coronary artery disease. Circulation **111**(25), 3481–3488 (2005)

[40] Tian, H., Occhipinti, E., Nassibi, A., Mandic, D.P.: Hearables: Heart rate variability from ear electrocardiogram and ear photoplethysmogram (ear-ECG and ear-PPG). Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 1–5 (2023)

[41] Kim, H.-G., Cheon, E.-J., Bai, D.-S., Lee, Y.H., Koo, B.-H.: Stress and heart rate variability: A meta-analysis and review of the literature. Psychiatry Investigation **15**(3), 235 (2018)