Atrial Fibrillation Detection from Ambulatory ECG with Accelerometry Contextualisation: A Semi-Supervised Learning Approach

Alex E. Voinas[1,3], Devender Kumar[2,4], Jan Smeddinck[2], Andreas Stochholm[3] and Sadasivan Puthusserypady[1]

***Abstract*— Atrial fibrillation (AF) is a common cardiac arrhythmia causing severe complications if left untreated. Due to its sporadic nature, early detection often requires longitudinal ambulatory electrocardiogram (ECG) screening. Recently, deep learning (DL) has gained prominence in analysing long-term ECG and automating AF detection. However, like any medical classification problem, obtaining diverse labelled ECG data for DL model training is expensive and time-consuming. This paper proposes a semi-supervised learning (SSL) based AF detection model employing a variational auto-encoder (VAE). It leverages varying amounts of labelled and unlabelled ECG data to optimise the AF detection performance on ambulatory ECG. As ambulatory contexts under free-living conditions influence ECG recordings, we incorporate context via accelerometry data and experiment with its influence on model performance. The proposed SSL model was trained on ECG data from 72,003 unique patients and can classify between sinus rhythms, AF, and other arrhythmias. Experimental results on unseen test dataset and the publicly available CACHET-CADB dataset clearly demonstrate the model's generalisability, achieving an accuracy of over 91% with just 20% of the training set being labelled. With extensive experiments, our study exhibits the ability of SSL to improve AF detection from ambulatory ECG using small amounts of labelled data.**

I. Introduction

Heart arrhythmias, abnormalities in the rhythm or rate of the heartbeat, are one of the leading causes of death worldwide. Among them, Atrial Fibrillation (AF) is one of the most prevalent. It increases the risk of getting a stroke by up to 4-5 times, hence if left untreated, it poses a significant threat to the subject's life expectancy. Further, AF is linked to an annual healthcare cost of €73–95 million in Denmark and an average of $26 billion in the USA [1], [2]. Electrocardiogram (ECG) is a non-invasive, cheap and easy-to-use technique for monitoring and the early detection of AF [3]. However, AF is often left undetected during the short-term in-hospital ECG recordings due to its paroxysmal nature i.e. appearing at random sporadic times [1]. Therefore, it is critical that long-term ambulatory monitoring is performed in the patients' natural settings, outside the clinic. The manual analysis of such long-term ECG for arrhythmia detection, involving visual examination, is a resource-consuming task and limits the widespread adoption of long-term monitoring [4], [5].

Over the past few decades, several computer-based methods have been introduced for automatically detecting AF from ECG recordings [6], [7]. Most of these are based on classic machine learning (ML) methods that require prior manual feature engineering by a domain specialist. Additionally, sometimes the extracted features have limited viability to generalise on unseen data and are prone to suffer from noise introduced in the ambulatory recordings [6], [8], [9].

In the recent past, the emergence of deep learning (DL) has impacted the field of ECG arrhythmia detection, and several end-to-end supervised DL methods for arrhythmia detection have been proposed [3], [5], [8], [10], [11]. DL is advantageous to conventional ML models as it does not require prior feature engineering [12].

Despite notable advancements in arrhythmia detection models through supervised DL methods, the broader implementation of these approaches faces several limitations. Firstly, supervised learning methods require large amounts of accurately labelled data for model training [4], [13]. Secondly, given health data privacy and security concerns, collecting and labelling such ECG data is time-consuming, expensive, and error-prone. Further, the models trained on limited available public ECG datasets face the issue of generalising on the unseen data from the free-living conditions [1].

To tackle the challenge of limited labelled data being available, a promising solution lies in utilising semi-supervised learning (SSL), which combines both supervised and unsupervised learning techniques [14]. SSL allows models to incorporate unlabelled data (available in abundance) into the supervised learning process, expanding learning capacity and performance. This integration of unlabelled data enables the model to extract valuable patterns, insights, and generalisations that may not be apparent when relying solely on labelled data [15].

Although the SSL methods have shown promising results in other domains, their utilisation for arrhythmias detection remains under-explored. The existing SSL based works on ECG primarily fall under two categories: (1) rhythm prediction [4], [13], [16], [17], [18], [19] , and (2) single ECG beat classification [20], [21]. Interestingly, the majority of these publications have primarily used publicly available ECG datasets, which are limited in size and diversity (regarding the number of patients) to fully exploit the power of SSL for

[1]A.E.Voinas and S.Puthusserypady are with the Department of Health Technology, Technical University of Denmark, Kgs. Lyngby 2800, Denmark {efsavo, sapu}@dtu.dk

[2]D.Kumar and J.Smeddinck are with the Ludwig Boltzmann Institute for Digital Health and Prevention, Lindhofstr. 22, 5020 Salzburg, Austria {devender.kumar, jan.smeddinck}@dhp.lbg.ac.at

[3]A.E.Voinas and A.Stochholm are with Cortrium ApS, Erik Husfeldts Vej 7, 2630 Taastrup, Denmark {av, as}@cortrium.com

[4] D.Kumar is also with the University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark deku@mmmi.sdu.dk

making truly generalisable models.

Another difficulty in the widespread operationalising of DL or other ML models for arrhythmia detection is that the ambulatory ECG analysis independent of the patient's context is challenging and prone to errors. These models do not consider the patient's characteristics and the physical state or free-living context (in which the ECG was collected) in the model building [1], [22]. Even in manual ECG analysis, when cardiologists find ECG strips non-conclusive, they usually look for increased context or knowledge about arrhythmia epidemiology to improve diagnostics [23]. In our prior work, [5], to incorporate context awareness in ECG classification, we used manual context heuristics derived from the accelerometer to improve the DL model's performance. However, manually driven context heuristics have their limitation in model generalisation.

In this paper, we present a semi-supervised classifier based on a modified variational autoencoder (VAE), that distinguishes between 3 different rhythm classes: Sinus Rhythms, AF/Flutter, and Others. Through extensive experiments on a large dataset from over 72,000 unique patients, we demonstrate the SSL's ability to leverage the usage of unlabelled data and achieve state-of-the-art classification performance with just limited labelled data.

We train our VAE ranging from the near fully supervised to fully self-supervised case, by varying the proportion of the labelled data in the training set from 99% down to 1%. To test the model's generalisability, we evaluate it on another public dataset CACHET-CADB where it showed an accuracy of over 91% with just 10% of labelled data. We also used the accelerometer (ACC) data (recorded alongside the ECG) to incorporate relevant traces of the patient's ambulatory context into the model. Although the manually extracted context heuristics can help the model (as shown in [5]), the effect of using the ACC-based context in an end-to-end SSL model for AF detection remains unexplored. In summary, this work makes the following two novel contributions:

- We present a state-of-the-art arrhythmia detection model that showcases SSL's effectiveness on a large ambulatory dataset recorded from over 72,000 patients, which underscores its impact in the domain.
- We investigate the effectiveness of ACC-driven context features in an end-to-end SSL-based arrhythmia classification model.

This research work has been conducted in collaboration with Cortrium ApS [24], which generously provided the primary training ECG dataset recorded from the extensive cohort of 72,003 subjects from 15 countries.

II. Materials and Methods

A. Dataset

The dataset used in this work is provided by Cortrium ApS, recorded using their ambulatory ECG monitor "C3+" [24]. Cortrium has obtained consent through the doctors using the C3+ Holter monitor and maintains the right to utilise the ECG data in pseudonymised form for research and development purposes. It consists of 229,913 7-second ECG strip examples, from 72,003 subjects in total, who wore the device in free-living conditions for over 24 hours long recording. Each 7-second ECG example has been manually reviewed and labelled by Cortrium's in-house clinical expert quality control team. All recordings consist of a 3-channel ECG signal, sampled at 256Hz [24]. The ECG recordings are accompanied by a "lead-off" indicator for the electrode attachment and readability quality. Additionally, the recordings also include 3-channel accelerometer (ACC) data sampled at the rate of 1/12th of the ECG signal i.e. 21.33Hz.

The Cortrium dataset comprises the following heart rhythm strip types: 66,028 Normal Sinus Rhythms (NSR), 42,802 Sinus Tachycardias (TachyN), 22,357 Sinus Bradycardias (BradyN), 28,143 Atrial Fibrillation examples (AF; in which Atrial Flutter is also included), 22,815 Supraventricular Runs (RunsS), 2,316 Ventricular Runs (RunsV), 15,099 Supraventricular Tachycardias (TachyS), 184 Ventricular Tachycardias (TachyV), 5,556 Supraventricular Bigeminies (BigeS), 6,366 Ventricular Bigeminies (BigeV), 6,163 Supraventricular Trigeminies (TrigeS), 6,551 Ventricular Trigeminies (TrigeV), 103 Idioventricular Rhythms (IVR) and 5,430 AV-block examples.

To evaluate the model's generalisability on a completely unseen dataset, recorded using a different ECG device and demographics, we also use the publicly available dataset CACHET-CADB [25]. It consists of 1602 10-second single-channel ECG examples including 3-channel ACC data, sampled at 1024Hz and 64Hz, respectively. The dataset comes from 24 subjects and includes 4 different ECG rhythm classes, with their breakdown being 747 AF, 615 NSR, 221 Noise, and 19 episodes of "Other" rhythm types [25].

B. Data reorganisation

Our proposed method solves a 3-class classification problem for Sinus Rhythms (SR), AF, and Other arrhythmias. Therefore, before training the model, we reorganise the dataset, into these three distinct classes keeping their representation equal in the dataset. The SR and "Other" are umbrella classes that include many sub-classes. The SR contains NSR, BradyN, and TachyN, while "Other" comprises a combination of the rhythm types RunsS, RunsV, TachyS, TachyV, BigeV, BigeS, TigeV, TrigeS, IVR, and AV-block. For the AF class, we take all the available 28,143 examples and we match the number of both the SR and Other classes to that number. The sub-classes within SR and Other classes were also balanced. For testing on the CACHET-CADB dataset, the noise class was excluded, and the remaining three sub-classes, namely NSR, AF, and 'Others,' were used as-is.

C. Data pre-processing

The ECG is kept raw and one of its three channels is randomly selected unless there is a "lead-off" indication, in which case the clean channel is automatically chosen. The ACC is converted to one channel, capturing the general body movement in space, by extracting its 3D magnitude using the

equation:

$$ACC_{mag} = \sqrt{ACC_x^2 + ACC_y^2 + ACC_z^2}. \quad (1)$$

The ACC is then upsampled to match the ECG's frequency and length. Finally, to distinguish the high power frequency zones contributing to any potential movement, its Power Spectral Density is calculated by squaring the magnitude of its Discrete Fourier Transform. Following this, the ECG and ACC signals are merged into a single data point with a sample length of 1792, (7-second strip; $7 \times 256 = 1792$). Finally, each data point is z-score normalised before being fed to the model. The same preprocessing is applied to the test dataset CACHET-CADB. A detailed block diagram of the preprocessing pipeline is shown in Figure 1.

D. Model Architecture

The SSL model developed in this work, for AF classification, is a modified variational autoencoder (VAE) [26]. Figure 1 shows the detailed architecture of our proposed model. The first neural network, after the Preprocessing step, named Encoder, maps a high dimensional input (x) into a dimensionally reduced Latent Space (z) which parametrises an approximated probability distribution expressed by latent variables. The Encoder works as an inference model expressed by $q_\phi(z|x)$ which is an approximation of the real distribution $p_\theta(z|x)$ [27]. Thereafter, sampling is performed from the Latent Space and the sampled output is passed through a second neural network, named Decoder, that maps it into a reconstruction ($\hat{x}$) of the original input. The Decoder works as a generative model expressed by $p_\theta(x|z)$ [27]. In our modified VAE version, the input ECG strip label (y) is also added to the set of latent variables as an additional "auxiliary" variable. This increases the model's variational flexibility, enabling it to be trained as a supervised model in parallel to being unsupervised, thereby achieving the semi-supervised learning capability [28]. The Classifier is an extra neural network accepting the output of the Encoder to perform a $K = 3$ class classification. It is defined by the expression $q_\phi(y|x)$. With the introduction of y, the Encoder-Classifier combo, mathematically expressed by $q_\phi(z|x,y)q_\phi(y|x)$, is represented by the Equations (2) and (3).

$$q_\phi(z|x,y) = \mathcal{N}\Big(z|\mu_\phi(x,y),\ diag\big(\sigma_\phi^2(x,y)\big)\Big), \quad (2)$$

$$q_\phi(y|x) = \mathrm{Cat}_3(y|\pi_{\phi_K}(x)). \quad (3)$$

In our implementation, the model's learned parameters of mean μ_ϕ and $\log(\sigma_\phi^2)$ (log-variance) describing the Gaussian distribution of $q_\phi(z|x,y)$ are represented by the Encoder network's last layer. They are split into two channels, one for each. Then, the Sampler will pick a z based on these parameters using the reparametrisation trick [29]. Our generative Decoder is defined as $p(z)p_\theta(x|z,y)$, the terms of which refer to the latent space approximated distribution and likelihood of reconstructing the input x correctly. Mathematically they are expressed by the Equations (4) and (5).

$$p(z) = \mathcal{N}(z|0,\mathrm{I}), \quad (4)$$

$$p_\theta(x|z,y) = \mathcal{N}(x|\hat{x},\ diag(0.025)). \quad (5)$$

Following the approach of Maaløe et al. [28], all our approximated distributions are Gaussians. Hence, $p(z)$ is a normal Gaussian and the reconstruction likelihood term $p(x|z,y)$ is a Gaussian with parameters $\mu_\theta = \hat{x}$ (reconstruction of the input) and $\sigma_\theta = 0.5$.

In our implementation, the Encoder consists of a 4-block ResNet architecture [30]. Each block consists of 3 convolutional layers and a residual connection, followed by a Max-pooling layer (kernel size=3, stride=3) reducing the input signal by a factor of 3. The blocks' output channels are of sizes 24, 32, 48 and 64, respectively.

The ResNet architecture is chosen due to its superior performance in many computer vision tasks and widespread use in supervised ECG signal classification [20], [30], [31]. At the tail of the Encoder, two Fully-Connected (FC) layers of size 1000 and 500 are applied. The model's latent space is described by 32-dimensional vectors, from which we pick 10 samples at a time. Likewise, the Decoder is a neural network of 3 FC layers with sizes 500, 1000, and 7168.

The last layer produces the reconstruction of the input as $\hat{x}$ and its log-var $\log({\sigma_{\hat{x}}}^2)$ with the dimension of size $2 \times 1792 \times 2 = 7168$, including one channel for the ECG and one for ACC.

The Classifier is a neural network of 3 FC layers with sizes of 500, 100, and 3, followed by a Softmax layer for outputting the class predictions. All the layers included in the network use the ReLU activation function, batch-normalisation, and dropout with a 1% probability.

E. Model optimisation and training

Based on variational inference, the VAE is optimised by maximising the Evidence Lower Bound (ELBO) function [26]. The likelihoods we want to maximise are $\log(p(x,y))$ and $\log(p(x))$, which cover the cases of labelled and unlabelled optimisation respectively [4].

The optimisable objective function (ELBO) for the labelled case is given by Equation (6).

$$\begin{aligned}\log p(x,y) &= \log \int_z p_\theta(x,y,z)dz \\ &\geq E_{q_\phi(z|x,y)}\left[\log \frac{p_\theta(x,y,z)}{q_\phi(z|x,y)}\right] \equiv -\mathcal{L}(x,y).\end{aligned} \quad (6)$$

In the case of unlabelled data, where y participates to the set of latent variables and is sampled from summing over the 3 classes, is represented by Equation (7).

$$\begin{aligned}\log p(x) &= \log \sum_y \int_z p_\theta(x,y,z)dz \\ &\geq E_{q_\phi(z,y|x)}\left[\log \frac{p_\theta(x,y,z)}{q_\phi(z,y|x)}\right] \equiv -\mathcal{U}(x).\end{aligned} \quad (7)$$

The ELBOs are composed of 2 terms. First, the KL-divergence, implying how much the real distribution $p_\theta(*|\mathrm{x})$ and its approximation $q_\phi(*|\mathrm{x})$ look-alike, which in the case of Gaussians is given by:

$$KL\big(q_\phi||p_\theta\big) = \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log((\sigma_j^2) - (\mu_j)^2 - (\sigma_j)^2\right), \quad (8)$$

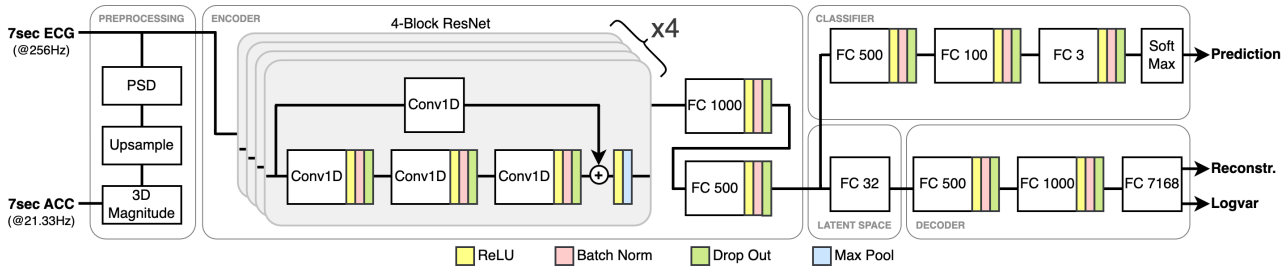


Fig. 1: The full diagram of the proposed SSL pipeline, including the data preprocessing and major parts of the VAE model.

where, μ and σ are functions of x and the variational parameters ϕ [26]. Second, the log-likelihood term, describing the probability that $\hat{x}$ is a correct reconstruction of x. It basically acts as the reconstruction loss. In the labelled case only, we introduce an explicit classification loss for adding the label-based y optimisation to our objective. We use the Multi-class Cross-Entropy loss, given by:

$$CE_3 = \sum_{l}^{K=3} y_{\hat{l},l} \log q_\phi(y_l|x_l), \tag{9}$$

where K is the number of classes (in our case $K = 3$), $y_{\hat{l},l}$ is the binary indicator if class label l is the correct classification for observation $\hat{y}$ and $q_\phi(y_l|x_l)$ is the probability of predicting the label y_l correctly. The objective expression for the labelled case is represented by Equation (10).

$$\mathcal{L}_l(x_l, y_l) = \mathcal{L}(x_l, y_l) - \alpha \cdot CE_3. \tag{10}$$

Finally, the full objective expression to optimise (loss function), combining both labelled and unlabelled cases, is given by:

$$\mathcal{J} = \sum_{(x_l,y_l)} \mathcal{L}_l(x_l, y_l) + \sum_{(x_u)} \mathcal{U}(x_u). \tag{11}$$

In all the experiments, our model is trained for 50 epochs with a batch size of 30. Therefore in one batch round, a set of 30 labelled and 30 unlabelled instances pass through the model separately to contribute to the optimisation of $\mathcal{J}$ (Eq. 11) before making a backpropagation step. To compensate for the number imbalance between the labelled and unlabelled batch groups (except for the 50% labelled - 50% unlabelled that are equal), at each epoch we cycle over the same batches of the smaller group until matching the number of batches of the bigger group. To make our model train better we introduced three warm-ups. Firstly, for the learning rate, which is initialised at 0.0001 at epoch 1. Then, at the epochs 5, 28, 40 and 45, it is divided by 2, 5, 2 and 2, respectively. Secondly, we gradually increase the effect of the classification loss (α term in Eq.10), starting from 0.125 and increasing with steps of 0.125 until reaching its maximum value of 0.5 at epoch 40 and thereafter remaining constant. Thirdly, in order to put more weight on the generative aspect of the model in the beginning, we use the KL Annealing method [32], letting only the reconstruction and classification losses have an effect in the first epochs of training. Hence, we delay the effect of the KL-divergence until epoch 25 and gradually increase it from epoch 26 onwards with steps of 0.001. The hyperparameters of the model were tuned using the random search method [33] and Adam optimiser [34].

F. Experiments

Our experiments are focused on showcasing the ability of the SSL model to train using a limited amount of labelled data. To achieve this, we carry out a range of experiments in which we alter the ratio of labelled data within the training set to the following percentages: 1%, 5%, 10%, 15%, 20%, 40%, 50%, 60%, 80% and 99%. It's important to note that when a specific percentage of the training data is utilised as labelled data, the remaining portion of the training data is utilised as unlabelled. This way, the whole spectrum ranging from nearly fully self-supervised (1% labelled) up to nearly fully supervised (99% labelled) is covered. Further, we also experiment with the effect of adding ambulatory context via ACC data. We compare the model's performance with and without ACC-based context for all the percentages of labelled and unlabelled data mentioned above.

G. Model validation

The model is validated using the hold-out method, where the dataset examples are divided into 90% training and 10% testing data. For each experiment, we perform random index selection, and in order to better keep track of the real-world scenario, smoothing out the effect a single index selection might have on our results, we repeat each experiment 15 times and take the average performance of each.

III. Results

A. Results on Cortrium test dataset

The detailed results of the model's performance on the test dataset (10% of the Cortrium dataset kept aside for testing) are shown in Table I and compared with other state-of-the-art methods in Table III. All the values are accompanied by their standard deviations since they are the mean result of 15 experiment repetitions. The column "% Labelled" depicts the percentage of the labelled data used in the training. The cumulative confusion matrix of the model's performance on 1% and 20% of labelled data on the Cortrium test dataset with ECG+ACC input are shown in Figures 2a and 2b, respectively. Figure 3a describes the Receiving Operating Characteristic (ROC) curve and Area Under the Curve (AUC) for each class for the model trained on 20% labelled data and ECG+ACC as inputs. For the model with the combination of ECG+ACC

TABLE I: Performance results (in %) from evaluating the VAE model on the main Test set (10% of Cortrium dataset). The column '% labelled' depicts the percentage of the labelled data used in the training.

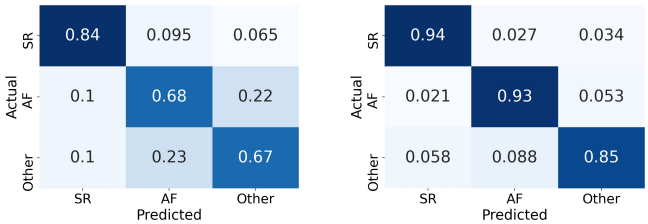
	ECG+ACC					ECG only				
%Labelled	**Accuracy-all**	**Precision-AF**	**Recall-AF**	**Specificity-AF**	**F1-AF**	**Accuracy-all**	**Precision-AF**	**Recall-AF**	**Specificity-AF**	**F1-AF**
1	72.83 ± 5.20	67.47 ± 5.66	67.56 ± 7.81	83.77 ± 2.82	67.46 ± 6.49	73.31 ± 4.13	67.91 ± 6.27	67.78 ± 7.92	84.07 ± 2.85	67.82 ± 7.01
5	87.15 ± 0.62	84.38 ± 1.69	89.12 ± 2.11	91.71 ± 1.2	86.65 ± 0.77	86.82 ± 1.80	84.27 ± 2.62	88.07 ± 3.08	91.77 ± 1.46	86.11 ± 2.53
10	89.32 ± 0.65	87.77 ± 1.01	90.31 ± 2.21	93.69 ± 0.68	89.0 ± 0.95	89.34 ± 0.98	88.03 ± 1.31	90.37 ± 1.3	93.84 ± 0.78	89.18 ± 0.84
15	90.23 ± 0.42	88.75 ± 1.38	91.63 ± 1.32	94.17 ± 0.88	90.15 ± 0.59	90.25 ± 0.51	88.83 ± 1.24	91.66 ± 1.32	94.22 ± 0.77	90.21 ± 0.63
20	90.64 ± 0.67	89.01 ± 1.58	92.62 ± 1.57	94.26 ± 1.00	90.76 ± 0.65	90.46 ± 1.08	88.41 ± 2.63	92.49 ± 1.18	93.88 ± 1.71	90.37 ± 1.32
40	91.46 ± 0.30	90.41 ± 1.04	92.82 ± 1.22	95.07 ± 0.64	91.59 ± 0.31	91.6 ± 0.34	90.22 ± 1.02	93.37 ± 1.02	94.93 ± 0.64	91.76 ± 0.44
50	91.57 ± 0.45	90.28 ± 0.77	93.32 ± 1.12	94.97 ± 0.47	91.77 ± 0.5	**92.03 ± 0.37**	**90.59 ± 1.37**	93.69 ± 0.80	**95.12 ± 0.83**	**92.1 ± 0.45**
60	**91.80 ± 0.55**	**90.92 ± 1.16**	93.00 ± 1.42	**95.34 ± 0.69**	**91.93 ± 0.61**	91.90 ± 0.28	90.39 ± 0.83	93.69 ± 0.66	95.01 ± 0.50	92.01 ± 0.34
80	91.36 ± 0.54	90.20 ± 2.00	93.00 ± 1.40	94.91 ± 1.22	91.55 ± 0.63	91.32 ± 0.47	89.23 ± 1.11	**93.74 ± 0.99**	94.33 ± 0.69	91.42 ± 0.52
99	91.58 ± 0.44	90.13 ± 1.64	**93.54 ± 1.31**	94.85 ± 1.02	91.79 ± 0.66	91.57 ± 0.46	89.92 ± 1.23	93.72 ± 1.01	94.73 ± 0.74	91.77 ± 0.54

TABLE II: Performance results (in %) from evaluating the VAE model on the CACHET-CADB dataset.

	ECG+ACC					ECG only				
%Labelled	**Accuracy-all**	**Precision-AF**	**Recall-AF**	**Specificity-AF**	**F1-AF**	**Accuracy-all**	**Precision-AF**	**Recall-AF**	**Specificity-AF**	**F1-AF**
1	60.97	69.38	70.68	63.25	70.03	63.43	73.37	60.11	74.29	66.08
5	81.68	79.78	96.12	71.29	87.19	74.22	83.72	91.57	79.02	87.47
10	91.31	98.12	90.63	97.95	94.22	87.98	91.72	91.97	90.22	91.84
15	89.36	95.63	93.71	94.95	94.66	90.73	94.36	94.11	93.38	94.24
20	92.83	98.47	94.51	98.26	96.45	91.96	95.28	94.65	94.48	94.96
40	92.83	97.09	93.84	96.69	95.44	92.69	**97.66**	94.78	**97.32**	**96.2**
50	92.76	97.75	93.17	97.48	95.41	92.9	97.5	93.98	97.16	95.71
60	**93.34**	**97.9**	93.71	**97.63**	**95.76**	92.03	97.38	94.38	97	95.85
80	92.69	95.46	**95.72**	94.64	95.59	92.69	96.97	94.38	96.53	95.66
99	92.61	94.95	**95.72**	94.01	95.33	**93.34**	95.35	**96.12**	94.48	95.73

TABLE III: Performance of State-of-the-art SSL methods on AF rhythm classification; and comparison with current work.

Paper	**Year**	**Dataset**	**#Ch**	**Recording**	**Input**	**Model**	**Target**	**Performance**
[4]	2021	MIT-BIH AF (AFDB)	2	In-hospital	Raw ECG	VAE (ResNet, FFN)	binary: AF, non-AF	Accuracy/Sensitivity/Specificity: 94.0%/98.9%/89.3% (for 1% labelled data)
[16]	2021	CinC2021 including 88,000 annotated ECGs	2	In-hospital	Raw ECG	Encoder-Decoder (CNN, Transformer, FFN)	multi-class: AF, AFLU and other arrhythmias	F1-score: 0.43
[13]	2022	1000 patients' 24hr ECGs	12	In-hospital	Feature-based	MMT: CNN-BLSTM-FFN	binary: AF, non-AF	Accuracy/Sensitivity/Specificity: 96.8%, 96.1%, 97.5%
[17]	2022	3,846,564 30sec ECGs and CinC2017	1	In-hospital	Feature-based	XGBoost, Manual	4-class: Sinus Rhythm, AF, Other, Noise/Artefact	F1-score: 89.0% (custom set) F1-score: 67.0% (CinC 2017)
[18]	2023	PTB-XL	12	In-hospital	Raw ECG	TCN	5-class: NORM, MI, STTC, CD, HYP	AUC/Accuracy/F1-score: 0.92 / 90.5% / 80.2%
[19]	2024	G12EC, PTB-XL, Ningbo Chapman-Shaoxing DB	12	In-hospital	Raw ECG	ECGMatch (Augment, Pseudo-Label, Correlation Align)	5-class: Abnormal, ST/T, CD, Normal, Other	Within-Dataset Macro-AUC: 0.93 Cross-Dataset Macro-AUC: 0.90 Mix-Dataset Macro-AUC: 0.84
This work	2025	230,038 7-sec ECGs from 72,003 subjects	3	Free-living	Raw ECG	VAE (ResNet, FFN)	3-class: SR, AF, Other	Accuracy/Sensiv-AF/Specif-AF: 90.6%, 92.6%, 94.2% (model using only 20% labelled data)
This work	2025	CACHET-CADB	1	Free-living	Raw ECG	VAE (ResNet, FFN)	3-class: SR, AF, Other	Accuracy/Sensiv-AF/Specif-AF: 92.8%, 94.5%, 98.2% (model using only 20% labelled data)



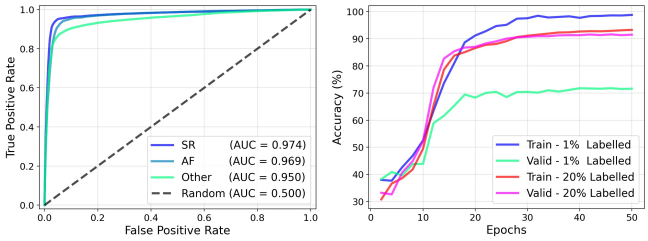
(a) 1% labelled (b) 20% labelled

Fig. 2: The Confusion Matrices for the 1% and 20% labelled ECG+ACC set of experiments, testing on the Cortrium test data (mean of 15 experiment repetitions).

inputs, initially, with just 1% of the labelled data, the model reaches a classification accuracy above 70%. At nearly 15%, it achieves an accuracy of over 90%, and in nearly fully supervised mode with 99% of the labelled data, it obtains an accuracy of 91.58 ± 0.44%. Interestingly, when only ECG is given as input, the model follows the same performance trajectory with the increase of labelled data.

B. Results on CACHET-CADB dataset

We tested the model's performance on the CACHET-CADB dataset to explore its generalisability. Table II shows the results on CACHET-CADB for the model trained on each percentage of the labelled data. Note that we randomly picked one out of the 15 trained model iterations for each percentage of labelled data. The performance on CACHET-CADB is also compared with other state-of-the-art methods in Table III. The model trained with 1% of labelled data achieved an overall accuracy of 60.97% and an F1-score of 74% for the AF class on ECG+ACC input. For ECG-only inputs, it has a slightly higher overall accuracy (63.43%) but a lower F1-score (66.08%) for the AF class. As the percentage of labelled data increases, the performance also drastically improves, and at 20% labelled data model gives accuracy and AF class F1-score of 92.83% and 96.45%, respectively.



(a) (b)

Fig. 3: (a) ROC-AUC curve for each class individually (One-vs-Rest method), for the 20% labelled ECG+ACC case, on Cortrium test data. (b) Training/Validation Accuracy curves for models trained by using 1% and 20% of labelled data and ECG+ACC as inputs.

The performance on CACHET-CADB reaches its peak on the model trained on 60% labelled data, and adding subsequent labelled data does not improve the performance. Similar to the Cortrium test dataset, on a nearly fully supervised (99% labelled data) instance, no significant difference exists between the model performance with ECG+ACC and ECG-only inputs on CACHET-CADB.

IV. Discussion

The experimental results of training the models on various percentages of labelled and unlabelled data in Table I demonstrate the SSL's ability to leverage unlabelled data and achieve state-of-the-art classification performance. With just 20% of labelled data, it was possible to achieve an overall accuracy of around 91%. The same pattern applies to the metrics of Precision, Recall, Specificity, and F1-score for the AF class, where with 20% labelled data, their plateau of performance is reached and remains until the 99% labelled case, with variability within the range of $\pm 1\%$.

To investigate the model's performance behaviour over the 50 epochs, we recorded the training and validation accuracy at each even-numbered epoch. Figure 3b depicts the training and validation curves of randomly picked examples from the 1% and 20% labelled data with ECG and accelerometer (ECG+ACC) as model input. It can be observed that when only 1% labelled data is used, the model overfits the training data as in the last epoch, the training accuracy reaches 98% while validation accuracy barely surpasses 71%. A plausible explanation for this behaviour could be that the stochasticity introduced by the small amount of labelled data constrains the model from learning the needed complex features between the three classes. When comparing to the 20% labelled case, it becomes evident that the model learns significantly better as both its training and validation curves follow each other closely over the whole of epoch progression. The model's performance improvement from 1% to 20% labeled data is statistically significant with $p < 0.001$ for both Accuracy and F1-scores.

Our proposed SSL model also exhibits substantial generalisability when applied to the unseen dataset CACHET-CADB. The performance on CACHET-CADB in Table II

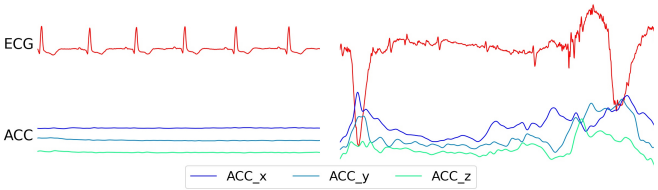


Fig. 4: ECG morphology and accelerometry data without ambulatory noises (left) and when distortion due to ambulatory context, such as body movement, is present (right).

follows a similar pattern to that of the evaluation on the main Cortrium test data (Table I). Although for models trained on the lower percentages (e.g., 1%, 5%, 10%) of labelled data, the generalisation on CACHET-CADB is slightly worse compared to the Cortrium test dataset. As explained earlier (Fig. 3b), this behaviour is expected on unseen datasets because the lower percentages of the labelled data models overfit the training set.

Further, it is interesting to note that all the experiments present some accuracy fluctuations after reaching their plateaus with nearly 20%-25% of labelled data. Ideally, the accuracy and F1-score should steadily increase while moving towards the fully supervised case. These fluctuations (past using 20% of the labelled data) might be because the generative aspect of the model (weight of KL-divergence) and classification loss get saturated after a certain percentage of labelled data and the generative aspect of the VAE gets restrained.

A. *Impact of accelerometer driven context on performance*

The investigation around the effectiveness of accelerometer-driven context features in an end-to-end SSL-based classification model (Tables I and II) revealed that including ACC features did not significantly impact the model's outcome ($p > 0.05$ when comparing all metrics for all labelled cases between ECG+ACC and ECG-only), especially in the higher percentages of labelled data. In our earlier work [5], it was observed that manually extracted context features from the accelerometer when used in post-processing heuristics, can effectively improve the supervised learning-based model's outcome. The ineffectiveness of the accelerometry features in the SSL model can have many plausible causes. Firstly, the Cortrium dataset used for the model training is relatively clean, and the noisy data was already pre-screened in the dataset. It is highly likely that accelerometry features are ineffective on non-noisy data, hence testing on more noisy data still remains an interesting area for further research. Typically even in manual ECG analysis, accelerometry data is more effective in differentiating if the distortion in ECG morphology is due to motion artefacts or heart physiology. In the case of clean and less ambulatory motion-affected ECG data, the accelerometry-driven features can be less effective, as the ECG morphological changes are directly linked to the heart's physiology (e.g. Fig. 4). Secondly, as both ECG and accelerometry data are directly fed into the model, they might be getting equal feature importance in the classification task.

V. Conclusion

This paper proposes a semi-supervised learning-based arrhythmia classifier using a modified VAE that can classify SR, AF and Other arrhythmias from ambulatory ECG recordings. Our state-of-the-art model showcases the effectiveness of semi-supervised learning on a large ECG dataset of 72,003 patients. It offers a potential solution to the problem of low availability of labelled ECG data–which, in turn, hampers the generalisability of supervised-learning-based ECG classification models. By experimenting with the range of different percentages of labelled data (1% to 99%) we showcase that using SSL with just 20% of labelled data, a classification performance (Accuracy 90.64 $\pm$ 0.67%) comparable to a near fully supervised model (Accuracy 91.58 $\pm$ 0.44% with 99% labelled data) can be achieved. Apart from the Cortrium test, the model trained on just 20% of labelled data also gave an accuracy of 92.83% and an AF F1-score of 96.4% on the CACHET-CADB dataset, indicating a good model generalisability on unseen datasets. Contrary to the hypothesis, our results indicate that adding accelerometry data did not impact the model's permanence, especially with higher percentages of labelled data. Nevertheless, exploring the impact of accelerometry-based context on the model's performance remains an interesting area of research, especially on ECG datasets with significant ambulatory noise.

References

[1] D. Kumar, "Designing and evaluating mobile health technology for ambulatory monitoring and diagnosis of heart arrhythmias," *DTU Health Technology, PhD Thesis*, 2021.

[2] S. P. Johnsen, L. W. Dalby, T. Täckström, J. Olsen, and A. Fraschke, "Cost of illness of atrial fibrillation: A nationwide study of societal impact," *Bmc Health Services Research*, vol. 17, no. 1, p. 714, 2017.

[3] E. R. Adams and A. Choi, "Using neural networks to predict cardiac arrhythmias," *Conference Proceedings - Ieee International Conference on Systems, Man and Cybernetics*, pp. 402–407, 2012.

[4] S. M. Rasmussen, M. E. K. Jensen, C. S. Meyhoff, E. K. Aasvang, and H. B. D. Sørensen, "Semi-supervised analysis of the electrocardiogram using deep generative models," *Engineering in Medicine and Biology Society (embc), 2020 42nd Annual International Conference of the Ieee*, vol. 2021, pp. 1124–1127, 2021.

[5] D. Kumar, A. Peimankar, K. Sharma, H. Domínguez, S. Puthusserypady, and J. E. Bardram, "Deepaware: A hybrid deep learning and context-aware heuristics-based model for atrial fibrillation detection," *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106899, 2022.

[6] S. M. P. Dinakarrao, A. Jantsch, and M. Shafique, "Computer-aided arrhythmia diagnosis with bio-signal processing: A survey of trends and techniques," *Acm Computing Surveys*, vol. 52, no. 2, p. a23, 2019.

[7] A. Petrenas, V. Marozas, and L. Sörnmo, "Low-complexity detection of atrial fibrillation in continuous long-term monitoring," *Computers in Biology and Medicine*, vol. 65, pp. 184–191, 2015.

[8] R. S. Andersen, A. Peimankar, and S. Puthusserypady, "A deep learning approach for real-time detection of atrial fibrillation," *Expert Systems With Applications*, vol. 115, pp. 465–473, 2019.

[9] M. Kumar, R. B. Pachori, and U. Rajendra Acharya, "Automated diagnosis of atrial fibrillation ecg signals using entropy features extracted from flexible analytic wavelet transform," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 3, pp. 564–573, 2018.

[10] U. R. Acharya, H. Fujita, O. S. Lih, Y. Hagiwara, J. H. Tan, and M. Adam, "Automated detection of arrhythmias using different intervals of tachycardia ecg segments with convolutional neural network," *Information Sciences*, vol. 405, pp. 81–90, 2017.

[11] H. Dang, M. Sun, G. Zhang, X. Qi, X. Zhou, and Q. Chang, "A novel deep arrhythmia-diagnosis network for atrial fibrillation classification using electrocardiogram signals," *Ieee Access*, vol. 7, no. 99, pp. 75 577–75 590, 2019.

[12] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[13] P. Zhang, Y. Chen, F. Lin, S. Wu, X. Yang, and Q. Li, "Semi-supervised learning for automatic atrial fibrillation detection in 24-hour holter monitoring," *Ieee Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3791–3801, 2022.

[14] M. F. A. Hady and F. Schwenker, "Semi-supervised learning," *Handbook on Neural Information Processing*, pp. 215–239, 2013.

[15] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine learning*, vol. 109, no. 2, pp. 373–440, 2020.

[16] R. Rodrigues and P. Couto, "Semi-supervised learning for ecg classification," *Computing in Cardiology*, vol. 2021-, p. 4 pp., 2021.

[17] F. J. Wesselius, M. S. van Schie, N. M. de Groot, and R. C. Hendriks, "An accurate and efficient method to train classifiers for atrial fibrillation detection in ecgs: Learning by asking better questions," *Computers in Biology and Medicine*, vol. 143, p. 105331, 2022.

[18] A. Ayad, M. Barhoush, M. Frei, B. Völker, and A. Schmeink, "An efficient and private ecg classification system using split and semi-supervised learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 9, pp. 4261–4272, 2023.

[19] R. Zhou, L. Lu, Z. Liu, T. Xiang, Z. Liang, D. A. Clifton, Y. Dong, and Y.-T. Zhang, "Semi-supervised learning for multi-label cardiovascular diseases prediction: A multi-dataset study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3305–3320, 2024.

[20] Z. Chen, Z. Lin, P. Wang, and M. Ding, "Negative-resnet: Noisy ambulatory electrocardiogram signal classification scheme," *Neural Computing and Applications*, vol. 33, pp. 8857–8869, 2021.

[21] H. Ziyang, Y. Shuaiying, Z. Jianhui, D. Bo, Y. Zhiyong, A. Adi, A. Fayadh, and A. Sara, "A novel myocardial infarction localization method using multi-branch densenet and spatial matching-based active semi-supervised learning," *Information Sciences*, vol. 606, pp. 649–668, 2022.

[22] D. Kumar, S. Puthusserypady, H. Dominguez, K. Sharma, and J. E. Bardram, "An investigation of the contextual distribution of false positives in a deep learning-based atrial fibrillation detection algorithm," *Expert Systems with Applications*, vol. 211, p. 118540, 2023.

[23] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature medicine*, vol. 25, no. 1, pp. 65–69, 2019.

[24] "C3+ holter monitor product overview and technical information," https://www.cortrium.com/product/, retrieved 2023-07-26.

[25] D. Kumar, S. Puthusserypady, H. Dominguez, K. Sharma, and J. E. Bardram, "Cachet-cadb: A contextualized ambulatory electrocardiography arrhythmia dataset," *Frontiers in Cardiovascular Medicine*, vol. 9, p. 893090, 2022.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *2nd International Conference on Learning Representations, Iclr 2014 - Conference Track Proceedings*, 2014.

[27] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[28] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther, "Auxiliary deep generative models," *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[29] C. Doersch, "Tutorial on variational autoencoders [arxiv]," *Arxiv*, p. 22 pp., 2016.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the Ieee Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-, pp. 770–778, 2016.

[31] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "Ecg heartbeat classification: A deep transferable representation," *Proceedings - 2018 Ieee International Conference on Healthcare Informatics, Ichi 2018*, pp. 443–444, 2018.

[32] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," 2016. [Online]. Available: https://arxiv.org/abs/1511.06349

[33] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.