Electrocardiogram Two-dimensional Motifs: A study Directed at Cardio Vascular Disease Classification

Hanadi Aldosari[1,4], Frans Coenen[1] , Gregory Y. H. Lip[2] Yalin Zheng[2,3]

[1] Department of Computer Science,University of Liverpool, Liverpool, UK
[2] Liverpool Centre for Cardiovascular Science, University of Liverpool and Liverpool Heart and Chest Hospital, Liverpool, UK
[3] Department of Eye and Vision Science, University of Liverpool, Liverpool, UK
{H.A.Aldosari, Coenen, Gregory.Lip, Yalin.Zheng}@liverpool.ac.uk
[4] College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia
hdosari@taibahu.edu.sa

Abstract. A process is described, using the concept of 2D motifs and 2D discords, to build classification models to classify Cardiovascular Disease using Electrocardiogram (ECG) data as the primary input. The motivation is that existing techniques typically first transform ECG data into a 1D signal (waveform) format and then extract a small number of features from this format for classification purposes. It is argued here that this transformation results in missing data, and that the consequent feature selection means that only a small part of the original ECG data is utilised. The approach proposed in this paper works directly with the image format, no transformation takes place. Instead, motifs and discords are extracted from the raw data and used as features in a homogeneous feature vector representation. The reported evaluation demonstrates that more effective classification results than that which can be achieved using the waveform format. The proposed 2D motif and discord extraction mechanism is fully described. The proposed process was evaluated using three distinct ECG data sets. A best accuracy of 85% was obtained, compared with a best accuracy of 68.48% using a comparable 1D waveform approach.

Keywords: 2D Motifs · 2D Discords · Cardiovascular Disease Classification · ECG Classification

1 Introduction

Cardiovascular Disease (CVD) is an umbrella term for a range of conditions that affect the heart and/or blood vessels, of which heart disease and stoke are perhaps the best known. Collectively, CVDs are the most common global cause of mortality, and the major contributor to reduced quality of life in the 21^{st} century [24]. According to the World Health Organisation (WHO) some 17.9 million people died from CVDs in 2019, representing 32% (approximately one third) of all global deaths [30]. The majority of these deaths (85%) were as a result of heart attacks or stroke. CVDs are most commonly caused by irregularities in the normal rhythm of the heart, the sinus rhythm. The sinus rhythm is between 60 and 100 beats per minute (bpm). A rate of less than 60bpm (sinus bradycardia) or above 100bpm (sinus tachycardia) is considered abnormal. The

standard tool for monitoring heart rate is the Electrocardiogram (ECG). ECGs are obtained using an ECG machine which detects and records the electrical signals produced by a patient's heart as it beats, using sensors attached to the patient's skin. Clinicians and cardiologists can then use the ECG data to assist in determining the presence, or otherwise, of CVD. This is achieved by examining individual heart cycles within the ECG trace in terms of what are referred to as the *P* wave, the *QRS* complex and the *T* wave. To speed up the ECG analysis process there has been significant interest in using the tools and techniques of machine learning. Especially the application of supervised learning to ECG data to build classification models of various kinds [11, 15, 21, 26].

Supervised learning requires labelled examples to which machine learning can be applied to generate a model that can then be used to label previously unseen examples. The labelled examples are usually divided into a training set and a test set. The first is used to "learn" the desired model, and the second is to evaluate the resulting model. The process of labelling the examples is often a time consuming and therefore a challenging task. A second challenge is how best to represent the examples so that machine learning can be applied. Most machine learning algorithms use a feature vector representation where the examples are represented using a numeric vector when each numeric value relates to a data attribute (feature, dimension). Generating such a representation is fairly straightforward if the data under consideration is in a numeric tabular format where each row represents a record and each column an attribute. This becomes much more challenging if our data is in the form of images, as in the case of ECG data.

ECG machines typically produced hard copy printouts. The first stage in the process of generating training and test data is thus to scan the paper format ECGs so that they are available in a digitised image format. The second stage is then to extract the ECG signal trace from the digitised images so that the data is in a waveform format. Once the transformation has taken place the next stage is to extract features from the waveform data so that a feature vector representation can be derived. Usually, the features identified are associated with the *P* wave, the *QRS* complex, and the *T* wave, used in the manual analysis of ECG data; examples can be found in [14, 18, 25, 31, 32]. The consequence, it is argued here, is that the resulting labelling (classification) of previously unseen examples is not as good as it might be because of: (i) the approximations used to generate the waveform format and (ii) the small number of features typically considered.

To address the above, in [1], a solution was presented founded on the use of motifs [2]. The solution moved away from the traditional idea of applying machine learning to a small number of features extracted from ECG data that had first been transformed into a 1D waveform format, by considering the ECG data in its entirety as an image. The idea presented was to extract 2D motifs directly from ECG image data and use these motifs as the attributes in a Homogeneous Feature Vector Representation (HFVR). In this context, a motif is a frequently repeating pattern which is considered to be indicative of a particular CVD label (class). In 1D a motif is a sub-sequence of points within a larger point (time) series. In 2D this is sub-matrix within a larger matrix of points (pixels). The concept of motifs, in the 1D context, is most frequently used in time series analysis [3, 20, 23, 38]. In the 2D context, motifs have been applied to image analysis

[4, 13]; although, with the exception of [1], not with respect to ECG data (at least to the best knowledge of the authors).

In [1] evidence was provided indicating that the use of 2D motifs for generating CVD classification models, using supervised learning, outperformed models generated using more traditional approaches. The evaluation was conducted using a subset of the Guangzhou Heart Study data set [9], a subset directed at Atrial Fibrillation (AF), a common form of CVD that is indicated by an irregular, and often unusually fast, heart rate (140 bpm) caused by the "twitching" of the top (atria) chambers of the heart. AF is the most common form of irregular heart beat. If untreated the presence of AF increases the risk of stroke and heart failure. However, a criticism of the work presented in [1] is that the AF versus no AF Guangzhou data set comprised only 120 records. The work presented in this paper revisits the work presented in [1] by re-analysing the claims made using a much more rigorous evaluation than was originally presented. Two additional stages have also been added to the proposed model in [1], for the cases of large numbers of motifs/discords being generated or when we have imbalanced input data. For the evaluation presented here three data sets were used: (i) the AF versus no AF Guangzhou data sets also used in [1], (ii) the entire Guangzhou Heart Study data set of 1172 records categorised as normal versus abnormal and (iii) the Liverpool Heart and Chest data set. The later is a recently acquired data set, curated by the authors, directed at AF with reoccurrence versus AF without reoccurrence (a much more challenging classification than in the case of the two other data sets considered).

In [1] a Support Vector Machine (SVM) classification model was used. A SVM classification model was also used with respect to the work presented in this paper. Partly so that fair comparisons with the work presented in [1] could be made, and partly because SVMs are frequently used with respect to reported work directed at more traditional CVD classification, see for example [33].

The rest of this paper is structured as follows. A review of previous work relevant to the work presented in this paper is given in Section 2. Section 3 then presents a formal definition of the 2D motif extraction problem (in the context of ECG data). A more extensive description of the approach to 2D motif feature extraction, and the utilisation of these motifs, than that presented in [1], is given in Section 4. Section 5 then provides a critical and comprehensive evaluation of the approach. The paper is completed, in Section 6, with a summary, some key conclusions and some suggested avenues for future work.

2 Related Work

Detection and classification of anomalies within ECG data has become a significant area of research in the context of CVD studies. The motivation is the observation that the manual interpretation of ECG data is time consuming, and requires prior knowledge and skills, knowledge and skills that are often in short supply. A range of Machine Learning (ML) and Deep Learning (DL) algorithms have therefore been proposed with the aim of addressing the challenge associated with the human interpretation of ECG data [14, 18, 25, 32].

As noted in the introduction to this paper, a particular challenge of applying ML and DL to ECG data is that the raw data is typically in a paper format. Thus the starting point for any form of classification model generation, using ML or DL, is the scanning (digitising) of the paper format data into a 2D image format. As also noted earlier, the practice is then to transform the 2D digitised ECG data into a 1D waveform format. There are a range of tools available to convert 2D digitised ECG data to a 1D waveform format [5, 6, 12, 16, 19, 22, 29, 34]. Using these tools the resulting wave forms, generated from digitized paper ECG data, can be in a variety of formats; both txt or xml are popular. Some of these tools provide additional functionality. Therefore the tools available can be divided into three groups according to the functionality that they provide: (i) Digitization + transformation (no additional functionality) (ii) digitizing + transformation + feature extraction, and (iii) digitizing + transformation + feature extraction + classification. Examples of the first can be found in [5, 6, 12, 29]. Example of the second can be found in [19, 22]. The idea here is to extract a small number of global characteristics from the ECG wave form data. As noted earlier, these characteristics are typically the amplitude and interval values of what are referred to as *P* wave, the *QRS* complex, and the *T* wave. The extracted characteristics can then be used to build a classification model. The last of the above tool groupings describes tools that incorporate classification model generation, not the case with respect to the previous two. Examples can be found in [16, 34]. In [16], morphological features were extracted to be used with two classification models, *k*NN coupled with Dynamic Time Warping (DTW) and Adaboost, to detect three different types of cardiovascular abnormalities. The digitisation tool used, in this case, was the same as that presented in [12]. In [34], the focus was on four specific types of waveform. The reported evaluation indicated that SVM model generation produced the best classification results.

An alternative to the waveform format, and that is explored in this paper, is to extract discriminatory features from the 2D scanned ECG image data without transformation to a 1D format, this can avoid the information loss associated with such transformations. The challenge is then the nature of the image features to be extracted. Classic approaches which involve the extraction of "low-level" features, such as colour or texture, are deemed to be ineffective for CVD disease classification [7]; More advanced feature extraction mechanisms are required. The solution proposed in [1] was to use 2D *motifs*. Motifs, as noted earlier, are repeating patterns found in data that can be used in tasks like clustering, classification, and anomaly detection. The idea of 2D motifs was first proposed in [4] and used in [13] for the purpose of classifying digital images featuring buildings, and images extracted from video news clippings, using a K-Nearest Neighbors (*k*NN) classification model. The work presented in this paper builds on the work presented in [3].

3 Problem Definition

This section provieds a formal problem definition for the work presented in this paper. The main goal is to generate a classification model that can be used to label previously unseen *digital ECG images* according to a given set of classes $C = \{c_1, c_2\}$. Each image I comprises a $n \times m$ pixel matrix such that p_{ij} is the pixel at row i and column j. The

input data D comprises a set of tuples of the from $\langle I, c \rangle$ where I is an ECG image and c_i is a class label taken from a set of classes C.

To generate the desired CVD classification model we aim to extract a set of features from each digital ECG image I. The idea proposed in this paper is that the most appropriate features to identify are 2D *motifs* and *discords*. A 2D motif M is a $p \times q$ sub-matrix of an image I, where $p < n$ and $q < m$, that occurs with maximal frequency. The intuition here is that because the sub-matrix occurs frequently it is likely to be a good discriminator of class. A motif set, $\mathbf{M} = \{M_1, M_2, \dots\}$, is a set of 2D motifs extracted from the images held in D, distributed according to class, one class per set of motifs. In other words, there is a one to one correspondence between the set $\mathbf{M}$ and the set of classes C, each subset $M_i \in \mathbf{M}$ is the set of motifs associated with the class $c_i \in \mathbf{C}$. Not all the motifs in $\mathbf{M}$ will be good discriminators of class, so it is necessary to prune $\mathbf{M}$. A two step process was adopted to achieve this, intra-class pruning to give the set M' and then inter-class pruning to give the set $\mathbf{M}''$.

A 2D discord S, in turn, is a $p \times q$ sub-matrix of an image I, of width p and height q, that occurs with minimal frequency (thus the opposite of a motif). The intuition here is that because the sub-matrix occurs infrequently it is likely to also be a good discriminator of class. A discord set, $\mathbf{S} = \{S_1, S_2, \dots\}$, is set of 2D discords extracted from the images held in D, again distributed according to class. As in the case of the motif set of sets $\mathbf{M}$, not all the discords in $\mathbf{S}$ are assumed to be good discriminators of class. Therefore, as in the case of motifs, we apply intra-class pruning to $\mathbf{S}$ to give S', and then inter-class pruning to give $\mathbf{S}''$. Further discussion concerning the intra- and inter-class pruning processes, with respect to the sets $\mathbf{M}$ and $\mathbf{S}$, is presented in Subsection 4.2.

4 Cardiovascular Disease Classification Model Generation

This section builds on the approach proposed in [1], adding two additional stages required when large numbers of motifs/discords are generated and/or when we have imbalanced input data. A schematic of the process is presented in Figure 1. From the figure it can be seen that the approach comprises seven stages:

1. Data cleaning.
2. 2D motif and discord extraction.
3. Feature selection.
4. Data augmentation.
5. Feature vector generation.
6. Classification Model Generation.
7. Classification Model Usgae.

Of these, feature selection and data augmentation are the two additional stages not originaly included in the process as first described in [1]. Detail concerning each of these five stages is presented in the following seven sub-sections, Sub-sections 4.1 to 4.7.

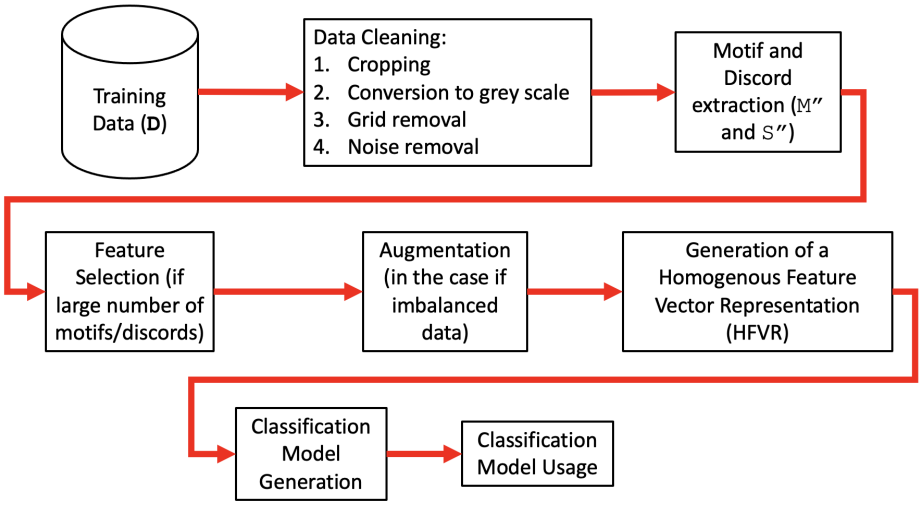


Fig. 1: Schematic of CVD Classification Model Generation Process

4.1 Data Cleaning (Stage 1)

The first stage in the CVD classification model generation process is the cleaning of the raw image data. The input to Stage 1 was a set of ECG images $\mathbf{D} = \{\langle I_1, c_1\rangle, \langle I_2, c_2\rangle\}$ where I_i is an ECG image and $c_i \in C$. From Figure 1 the Data Cleaning stage comprises four sub-processes: (i) cropping, (ii) conversion to gray scale, (iii) grid removal and (iv) noise removal. Each of the four data cleaning sub-processes is considered in the remainder of this section.

Cropping: Scanned ECG images often include spurious information around the edges of the scan. The first process was therefore to crop the image so that only the ECG signals were retained.

Conversion to Gray Scale: The cropped RGB image was then converted to a gray-scale intensity image.

Grid Removal: The third sub-processes was directed at removing all spurious data in the gray-scale ECG data, particularly the background grid which is a frequent feature of ECG digital images. This was achieved using the application of a "binarization" operation designed so that pixels related to the ECG traces were allocated the value 255 (white) and the rest of the image pixels the value 0 (black). The desired effect was that the graphical grid, and the majority of spurious data points and noise, would all be encoded as black pixels. The challenge was deciding the value of the binarisation threshold to be applied to the gray-scale image. To decide the nature of this threshold, histograms for selected ECG image files were generated. From these histograms, it was found that the background (high intensity) gray scale values were in the range 150 to 255, the threshold value was therefore set at 150. Thus, the proposed binarisation process assigned a value of 0 to each gray scale pixel whose value was greater than the 150 threshold, and a value of 255 otherwise (Equation 1).

Noise Removal: The anticipation was that some spurious small patches of white pixels (white noise) would be retained after the application of the binarization. To remove this white noise a *morphological erosion operation* was applied whereby the pixels in the boundary of white objects were removed. This also had the effect of reducing the thickness of the ECG traces. Thus, on completion of the erosion operation a *morphological dilation operation* was applied to add pixels back to the boundaries of the retained white objects, namely the ECG trace.

$$binary(x,y) = \begin{cases} 0 & \text{if } grayscale(x,y) > threshold \\ 255 & \text{otherwise} \end{cases} \quad (1)$$

4.2 Motif and Discord Extraction (Stage 2)

The second stage in the overall process was 2D Motif and Discord extraction (discovery). The pseudo-code for the top-level motif and discord extraction algorithm is given in Algorithm 1. Note that this algotithm is similar to that given in [1]. The inputs are: (i) the data set $\mathbf{D} = \{\langle I_1, c_1\rangle, \langle I_2, c_2\rangle, \dots\}$ where each image I_i has been pre-processing during Stage 1 of the process (see above), (ii) the set of classes C, (iii) the width p and height q of the motifs and discords to be retrieved, (iv) a pre-specified similarity threshold σ used to determine whether two pixel sub-matrices are the same or not, and (v) k the number of motifs and discords to be selected. The output was a set of pruned motifs and a set of pruned discords, $\mathbf{M}''$ and $\mathbf{S}''$, to be used in the following stages. The algorithm commences (lines 3 to 6) by segmenting the set D into a set of subsets $\mathbf{D} = \{D_1, \dots D_{|C|}\}$ where each subset is associated with a class in C. Note that for the evaluation presented in Section 5, $|C| = 2$ was used, hence $\mathbf{D} = \{D_1, D_2\}$. Two sets are then defined, lines 7 and 8 to hold identified motifs and discords, the sets $\mathbf{M}$ and $\mathbf{S}$.

The set $\mathbf{D}$ is then processed to identify the motifs and discords held in the images associated with each class (lines 9 to 18). This involves calls to a number of sub-processes which will be discussed in further detail later in this sub-section. The output is the set $\mathbf{M} = \{M_1, M_2, \dots\}$ and the set $\mathbf{S} = \{S_1, S_2, \dots\}$; where M_i is the set of motifs associated with class $c_i \in C$, and S_i is the set of discords associated with class $c_i \in C$. As noted earlier, for the evaluation presented in Section 5, $|C| = 2$ was used, hence $\mathbf{M} = \{M_1, M_2\}$, and $\mathbf{S} = \{S_1, S_2\}$. Note that the proposed approach may result in the same motif being identified in several images, thus $\mathbf{M}$ and $\mathbf{S}$ are likely to contain repeat occurrences of motifs and discords. The intuition here for them being retained was that they would be given more significance with respect to the generation of the intended prediction model; conceptually they would be given a higher "weighting".

The sets $\mathbf{M} = \{M_1, M_2, \dots\}$ and $\mathbf{S} = \{S_1, S_2, \dots\}$ are likely to hold some motifs and discords that are unique to only one image. It was anticipated, that these would not be good discriminators of class, hence, for each set of motifs $M_i \in \mathbf{M}$ associated with a particular class $c_i \in C$, and each set set of discords $S_i \in \mathbf{S}$ associated with a class $c_i \in C$, intra-class pruning was applied (line 19) and unique motifs and discords removed, The retained motifs and discords were held in the sets $\mathbf{M}' = \{M'_1, M'_2, \dots\}$ and $\mathbf{S}' = \{S'_1, S'_2, \dots\}$ respectively (line 16 in Algorithm 1).

The last sub-processes in Algorithm 1, line 19, was to conduct inter-class pruning. The removal of motifs and discords, from $\mathbf{M}'$ and $\mathbf{S}'$ respectively that were associated with more than one class and hence deemed to not be useful for distinguishing between classes. The retained motifs and discords were held in the sets $\mathbf{M}'' = \{m_1, m_2, \dots\}$, and a set of discords $\mathbf{S}'' = \{s_1, s_2, \dots\}$, that were considered to be good discriminators of class.

Algorithm 1 2D Motif and Discord Extraction [1]

1: **Input** D, C, p, q, σ, k
2: **Output** $\mathbf{M}'', \mathbf{S}''$
3: $\mathbf{D} = \{D_1 \dots D_{|C|}$ where $\forall D_i \in \mathbf{D}, D_i = \emptyset$
4: **for** $\forall \langle I_i, c_i \rangle \in D$ **do**
5: $D_j = D_j \cup I_i$ where $j = i$
6: **end for**
7: $\mathbf{M} = \emptyset$ ▷ Define the empty set **M** to hold extracted motifs
8: $\mathbf{S} = \emptyset$ ▷ Define the empty set **S** to hold extracted motifs
9: **for** $\forall D_i \in \mathbf{D}$ **do**
10: **for** $\forall I_j \in D_i$ **do**
11: $\chi_i \leftarrow$ genSubMatrices(I_j, p, q) ▷ Algorithm 2
12: $DM_i \leftarrow$ getCandidate2DmotifsAndDiscords(χ, σ) ▷ Algorithm 3
13: $M_i, S_i \leftarrow$ topK_2DmotifsAndDiscords(DM_i, k) ▷ Algorithm4
14: $\mathbf{M} \leftarrow \mathbf{M} \cup M_i$ ▷ Add M_i to the set **M**
15: $\mathbf{S} \leftarrow \mathbf{S} \cup S_i$ ▷ Add S_i to the set **S**
16: **end for**
17: **end for**
18: $\mathbf{M}', \mathbf{S}' \leftarrow$ intraClassPruning$(\mathbf{M}, \mathbf{S}, \sigma)$ ▷ Algorithm 5
19: $\mathbf{M}'', \mathbf{S}'' \leftarrow$ interClassPruning$(\mathbf{M}', \mathbf{S}', \sigma)$ ▷ Algorithm 6
20: **return** $\mathbf{M}'', \mathbf{S}''$

From the pseudo code given in Algorithm 1 it can be seen that the proposed 2D motif and discord extraction process comprises five sub-processes: (i) Generate sub-matrices, (ii) Generate candidate 2D motifs and discords, (iii) Get Top k 2D motifs and select discords, (iv) Intra-class pruning and (v) Inter-class pruning. Each of these is theefore discussed in further detail below.

Sub-matrix generation The pseudo code for the sub-matrix generation sub-process is given in Algorithm 2 (the algorithm is similar to that presented in [1]). The inputs are: (i) a pre-processed ECG image I associated with a particular class, and (ii) the desired sub-matrix window width d and height q. The sub-matrix window is slid over the image I pixel by pixel. The output is a set of sub-matrices, $\chi = \{Sub_1, Sub_2, \dots\}$ held within the image I. The algorithm commences, line 2, by defining the empty set χ in which to hold the extracted sub-matrices. Then, lines 3 to 7, the $p \times q$ sub-matrices in I are defined. We are only interested in sub-matrices that contain the ECG trace. Sub-matrices located at the edge of the image tended to be poor discriminators of class. Thus, sub-matrices that feature only black pixels and those located at the edge of the input image I were

not selected for inclusion in χ. This is tested for in line 4 of the algorithm. At the end of the process χ is returned (line 8). Note that if there are only "black" images in I, the set χ would be empty, although this would be an unlikely occurrence, and indicative of a faulty ECG input image.

Algorithm 2 Generate Sub-Matrices [1]

1: **Input** I, p, q
2: $\chi = \emptyset$ ▷ Define the empty set χ to hold pixel sub-matrices
3: **for** $\forall sub_i$ of size $p \times q \in I$ **do**
4: **if** $Sub_i \neq$ black and Sub_i not located on the edge of I **then**
5: $\chi = \chi \cup Sub_i$
6: **end if**
7: **end for**
8: **Return** χ

Candidate 2D motifs and Discords The pseudo code for generating candidate 2D motifs/discords is given in Algorithm 3, a similar algorithm was presented in [1]. The inputs are: (i) the set χ, associated with class i, of $p \times q$ sub-matrices generated using Algorithm 2, and (ii) the similarity threshold σ. The algorithm returns a set of candidate motifs/discords of the form $MD = \{\langle sub_1, count_1\rangle, \langle sub_2, count_2\rangle, \dots\}$ where $sub_i \in \chi$ and *count* is the corresponding occurrence count. The algorithm starts, line 3, by defining the empty set MD. The algorithm then processes each sub-matrix sub_i in χ (lines 4 to 13). First a counter, $count_i$, is defined and set to 0 (line 5), and $\langle sub_i, count_i\rangle$ added to the set MD (line 6). Sub-matrix sub_i is then compared to every other sub-matrix sub_j in χ, whenever a similarity between sub_i and sub_j is identified the count for sub_i is incremented by one and sub_j removed from χ (so that the same sub-matrix is not counted again later in the process). The similarity between the sub-matrices, sub_i and sub_j, is determined by calculating the Euclidean distance between the two matrices using Equation 2 given below. Euclidean distance measurement is frequently used for 1D motif similarity checking [35], and therefore was deemed to be appropriate for 2D similarity checking. The calculated Euclidean distance is then compared using the threshold σ, if the result is less than or equal to σ, sub_i and sub_j are deemed to be similar. The returned set MD will hold both candidate motifs and discords with respect to the input image (which will be associated with a particular class $c_i \in C$).

$$dist\,(sub_i, sub_j) = \sqrt{\sum_{h=1}^{h=(p\times q)} \left(md_{i_h} - md_{j_h}\right)^2} \tag{2}$$

Top K 2D Motifs and Discords Once a set of candidate motifs and discords for an image I associated with a class c_i, the set MD_i, has been identified, the next stage is to identify individual motifs and discords. Motifs will be the candidates associated with

Algorithm 3 Candidate 2D Motifs and Discords [1]

1: **Input** χ, σ
2: **Output** MD_i
3: $MD \leftarrow \emptyset$ ▷ Define the empty set MD to hold extracted motifs
4: **for** $\forall sub_i \in \chi$ **do**
5: $count_i \leftarrow 0$
6: $MD \leftarrow MD \cup \langle sub_i, count_i \rangle$
7: **for** $\forall sub_j \in \chi$, $j \neq i$ **do**
8: **if** dist $(sub_i, sub_j) \leq \sigma$ **then**
9: $count_i = count_i + 1$
10: $\chi \leftarrow \chi$ with sub_j removed
11: **end if**
12: **end for**
13: **end for**
14: **Return** MD

the highest counts, and discords with a count of one. The candidates in MD_i were thus ordered according to the associated frequency count and the top k were considered to be motifs, and those with a count equal to 1 to be discords. The pseudo code for achieving this is given in Algorithm 4; the algorithm is similar to that presented in [1]. The inputs are: (i) the set of candidate motifs and discords for an image i, the set $MD = \{\langle md_1, count_1 \rangle, \langle md_2, count_2 \rangle, \dots\}$ as generated using Algorithm 3, and (ii) the threshold k. The algorithm proceeds by first ordering the candidate motifs in MD according to their occurrence count (line 3). The top k are then selected as the chosen motifs and placed in M (line 4). Any candidate motifs with a count of 1 are deemed to be discords and placed in S (line 5). The sets $M = \{m_1, m_2, \dots\}$ and $S = \{s_1, s_2, \dots\}$ are then returned (line 6).

Algorithm 4 topK_2DmotifsAndDiscords [1]

1: **input** MD, k
2: **output** M, S
3: $MD_i \leftarrow MD$ sorted in descending order
4: $M \leftarrow$ top k candidates in MD_i
5: $S \leftarrow$ candidates in MD_i with a count of 1
6: **Return** M, S

Intra-Class Pruning We are interested in motifs and discords that are good discriminators of class. We are therefore not interested in motifs and discords that only appear in one image. Recall that M_i is the set of motifs associated with the class c_i, and that S_i is the set of discords associated with the class c_i. Thus, we wish to remove motifs and discords, from the sets $\mathbf{M} = \{M_1, M_2, \dots\}$ and $\mathbf{S} = \{S_1, S_2, \dots\}$ respectively, that appear in only one image (intra-class pruning). The sub-process for achieving this is

shown in Algorithm 5 (an identical algorithm was presented in [1]). The inputs are: (i) the set $\mathbf{M} = \{M_1, \ldots, M_{|C|}\}$, (ii) the set $\mathbf{S} = \{S_1, \ldots, S_{|C|}\}$ and (iii) a similarity threshold σ. The algorithm commences (lines 3 and 4) by declaring the empty sets $\mathbf{M}'$ and $\mathbf{S}'$ to hold the identified sets of motifs and discords; individual sets for individual classes. The set $\mathbf{M}$ is processed first, lines 5 to 11. For each motif m_j in the set $M_i \in \mathbf{M}$ (the set of motifs associated with class $c_i \in C$), if m_j does nor appear anywhere else in M_1 the motif is discarded, otherwise it is added to M'_i. A similar process is followed for the set $\mathbf{S}$, lines 12 to 18. At the end of the process the sets $\mathbf{M}'$ and $\mathbf{S}'$ will be returned. Note that it might be the case that the sets $\mathbf{M}'$ and $\mathbf{S}'$ are empty. Note also that determining whether a motif appears only in a single image requires a similarity comparison with the motifs for all the other images associated with the current class. This requires the similarity threshold σ. This is therefore a computationaly expensive task.

Algorithm 5 Intra-class pruning [1]

```
 1: input M, S, σ
 2: output M', S'
 3: M' ← {M'_1 ... M'_|C|} where ∀M_i ∈ M', M_i = ∅        ▷ Define the empty set M'
 4: S' ← {S'_1 ... S'_|C|} where ∀S_i ∈ S', S_i = ∅        ▷ Define the empty set S'
 5: for ∀M_i ∈ M do
 6:     for ∀m_j ∈ M_i do
 7:         if m_j appears in more than one image in M_i then
 8:             M'_i ← M'_j ∪ m_j
 9:         end if
10:     end for
11: end for
12: for ∀S_i ∈ S do
13:     for ∀s_j ∈ S_i do
14:         if s_j appears in more than one image in S_i then
15:             S'_i ← S'_i ∪ m_j
16:         end if
17:     end for
18: end for
19: Return M', S'
```

Inter-Class Pruning The last step in Stage 2 is to remove motifs and discords from M'_1 and S'_1 that are not good discriminators of class. In other words, motifs, and discords that associated with more than one class. The pseudo code for the inter-class pruning is given in Algorithm 6; the pseudo code is the same as that presented in [1]. The inputs are the sets $\mathbf{M}' = \{M'_1, M'_2, \ldots\}$ and $\mathbf{S}' = \{S'_1, S'_2, \ldots\}$ from the previous sub-process, and the similarity threshold σ. The outputs are the sets $\mathbf{M}'' = \{M''_1, M''_2, \ldots\}$, and $\mathbf{S}\prime\prime = \{S''_1, S''_2, \dot{\}}$, where M''_i is a motif and S''_i is a discord. The algorithm commences by declaring the sets $\mathbf{M}''$ and $\mathbf{S}''$ to hold the "double" pruned sets of motifs and discords. The set $\mathbf{M}'$ is processing first (lines 5 to 11), and the set $\mathbf{S}'$ second (lines 12 to 18). Line

7 states that if the the motif m'_j does not appear in the set of motifs associated with some other class, then m'_j should be added to $\mathbf{M}''$. Line 14 should be interpreted in a similar manner but with respect to discords. On completion, line 19, $\mathbf{M}''$, and $\mathbf{S}''$ are returned. To determine whether a motif or discord appears in the context of another class again requires similarity checking, which again entails the threshold σ to determine whether two motifs (discords) are the same or not.

Algorithm 6 Inter-class pruning [1]

```
1:  input M', S', σ
2:  output M'', S''
3:  M'' ← ∅                          ▷ Define the empty set M'' to hold double pruned motifs
4:  S'' ← ∅                          ▷ Define the empty set S'' to hold dooble pruned discords
5:  for ∀M'_i ∈ M' do
6:      for ∀m'_j ∈ M'_i do
7:          if ∀M'_k ∈ M, k ≠ i, m_j ∉ M'_k then
8:              M'' ← M'' ∪ m_j
9:          end if
10:     end for
11: end for
12: for ∀S'_i ∈ S' do
13:     for ∀s'_j ∈ S'_i do
14:         if ∀S'_k ∈ S, k ≠ i, S_j ∉ S'_k then
15:             S'' ← S'' ∪ s_j
16:         end if
17:     end for
18: end for
19: return M'', S''
```

4.3 Feature Selection (Stage 3)

The reported evaluation presented in [1] considered a single small data set. The more extensive evaluation conducted with respect to the work presented here (see Section 5) revealed that if the number of extracted motifs or discords exceeded 3,000 overfitting resulted. One solution might have been to reduce the value of the k parameter, the number of motifs extracted from an image. However, the work presented in [1] had demonstrated that $k = 5$ produced the best results. The adopted solution was therefore to include an additional stage in the overall process, Stage 3, that was invoked should the situation arise where more than 3,000 motifs were identified. The idea was to use a Dimensionality Reduction (DR) technique to reduce the number of features while attempting to keep as much of the variation in the original features set as possible [37]. There are many DR algorithms available for this purpose, for the work presented in this paper three methods were considered:

Principal Components Analysis (PCA). PCA operates using by performing a linear combination of the set of features. The combination was conducted in a given data

set so as to create a smaller set of features, in such a way as to capture as much information as possible in the smallest number of features. The resulting features are referred to as "Principal Components".

Singular Value Decomposition (SVD). SVD decomposes the original features by using the concepts of Eigenvalues and Eigenvectors into three constituent matrices to remove redundant features.

T-distributed Stochastic Neighbour Embedding (T-SNE). T-SME reduces the number of features by combining them into two or three new features. In a high dimensional space, the probability similarity of points is calculated. Consequently similar points are assigned a high probability, and dissimilar points are assigned a lower probability. Then, nearby points in the high-dimensional space are mapped to the nearest points in the low-dimensional space so as to achieve dimensionality (feature) reduction.

4.4 Data Augmentation (Stage 4)

As noted above, the experiment reported in [1] focused on a single, relatively, small data set (120 records). This data set also offered the advantage that it was balanced (an equal number of examples for each class considered). In practice balanced training data is unusual. This is often the case in the context of binary classification where there tends to be more examples of "normal" cases than "abnormal" cases. To address this issue, with respect to the work presented here, an *oversampling* technique was used to augment the minority class. In "classic" oversampling the minority data is simply duplicated. However, a criticism of this approach is that it will not add any new information, only existing information. Thus, the Synthetic Minority Oversampling Technique (SMOTE) technique [8] was adopted, a technique that can be used to synthesize new examples from existing examples. For the work presented in this paper three different SMOTE variations were considered:

The original SMOTE, which operates by first selecting random records from the minority class and finding the k-nearest neighbours to these records to create "clusters". Additional synthetic records are then created using these clusters.

Support Vector Machine SMOTE (SVM-SMOTE), which is similar to the original SMOTE but instead of using the K-nearest neighbours technique, a SVM model is used.

Adaptive Synthetic (ADASYN) SMOTE, which operates by considering the data density of the minority class and generating new examples in the less dense "areas".

4.5 Feature Vector Generation (Stage 5)

The fifth stage in the proposed approach (see Figure 1) was the generation the desired Homogeneous Feature Vectors Representation (HFVR) $H = \{V_1, V_2, \dots\}$. The idea here was that the HVRR, comprised of motifs and discords, would also allow for the addition of other features. In the evaluation presented later in this paper experiments are reported where clinical data were added. Each $V_i \in H$ is of the form $\{v_1, v_2, \dots, c\}$ where v_i is a numerical value, for example an occurrence count of a motif in $\mathbf{M}''$ or a discord in $\mathbf{S}''$,

in an ECG scanned image I_i. The final element, c, is a class label taken from a set of classes C. A previously unseen record will have a null value for the variable c as this is the value we wish to predict.

4.6 Classification Model Generation (Stage 6)

Once a suitable set of feature vectors have been generated the final stage was to generate the desired CVD classification model. The feature vector representation lends it self to many classification model generators (this was why this representation was selected with respect to the work presented here. In [33] and [34], SVM model generation was adopted for CVD classification. SVM model generation was also adopted in [1]. A SVM classification model was therefore also used with respect to the work presented in this paper.

4.7 Classification Model Usage (Stage 7)

Once the desired CVD classification model had been generated it could be applied to new data. In most cases this would be straight forward. However, in some cases, we may have more than one ECG image per patient, thus *multiple learning classification*. This was a feature of one of the evaluation data sets used for evaluation purposes (as reported on in Section 5). Thus some kind of "conflict resolution" was required where contradictory CVD classifications were produced. Three alternative options were considered on how to deal with this situation.

Averaging. Average the motif counts for each patient when identifying the motifs to be used, regardless of the number of ECG images considered, and used

Voting. Produce multiple classifications, one for each ECG image associated with a patient, and select the class that occurs the most frequently. In the event of a tie-break situation choose the class with the most serious consequences (in other words err on the side of caution).

Using only one image input. Thus avoiding the problem all together. In this case, the most recent ECG image was selected.

5 Evaluation

The extensive evaluation (more extensive than that discussed in [1]) of the CVD classification model generation mechanism is reported in this section. For the evaluation three data sets were used: (i) the subset of the Guangzhou Heart Study data set [9] concerned with AF which was also used in [1], (ii) the Guangzhou Heart Study data set in its entirety and (iii) a data set provided by the Liverpool Heart and Chest Hospital (LHCH). More detail concerning these data sets is provided in Sub-section 5.1. A SVM classification model was used with respect to all the experiments reported here, with Grid Search to choose the best parameters (C, gamma, and kernel). The evaluation metrics used were: accuracy, precision, recall, F1 score and AUC. Repeated Ten-fold cross-validation was used throughout. The Friedman Test was used to determine whether

or not there was a statistically significant difference between the performance. Where a statistically significant difference was identified, the Nemenyi post-hoc test was applied to identify the distinctions between the performances of the mechanisms considered. In [1] the evaluation objectives were focused on identifying the appropriate values for the parameters σ, k, p, and q. From [1] the most appropriate values were found to be:

- σ: The similarity threshold used to compare two motifs (the maximum distance between two motifs) $\sigma = 0.2$
- k: The number of most frequent candidate motifs to be selected from each image, $k = 5$
- p: The pixel matrix row size, $p = 30$
- q: The pixel matrix column size, $q = 90$

These were thus the values adopted with respect to the evaluation presented in this paper. The objectives of the evaluation were:

1. To identify the most appropriate feature selection (dimensionality reduction) and data augmentation techniques.
2. To identify the most appropriate conflict resolution technique where we have a "multiple learning classification" issue.
3. To compare the operation of the proposed approach when the motif/discord set was augmented in various ways with additional data.
4. To compare the operation of the proposed approach with a "traditional" 1D waveform approach.

Each of these objectives is discussed in further detail in the following four sub-sections, Sub-sections 5.2, 5.3, 5.4 and 5.5.

5.1 Data Sets

As noted in the introduction to this section three data sets were used for the evaluation presented here:

1. Guangzhou Atrial Fibrillation (GAF)
2. Guangzhou Heart Study (GHS)
3. Liverpool Heart and Chest Hospital (LHCH)

Some statistics concerning these data sets are given in Table 1. GAF and GHS, the first two data sets listed, were extracted from the Guangzhou Heart Study data set [9]. This comprised 1172 patients; each patient was associated with a 12-leads ECG scanned image and patient attributes, including age and gender. Each patient record had been labeled according to arrhythmia type, either sinus arrhythmia (normal) or abnormal. The abnormal category included: (i) Atrial Fibrillation (AF) and Flutter (AFL), (ii) Premature ventricular contractions, (iii) Premature atrial contractions , (iv) ventricular tachycardia, (v) Wolff-Parkinson-White syndrome, (vi) pacing rhythm and (vii) borderline rhythm. Each image was stored using JPEG compression with a resolution of 300 dpi (dots per inch).

Table 1: Statistics Concerning Evaluation Data Sets

Data	c_1				c_2			
Set	Label	# Rec.	# Male	# Female	Lable.	# Rec.	# Male	# Female
GAF	AF	60	32	28	Not AF	60	22	38
GHS	Normal	878	283	595	Abnormal	294	116	178
LHCH	No recurrence	639	428	211	Recurrence	270	182	88

The GAF data set comprised a subset of Guangzhou Heart Study data set that featured only two labels, Atrial Fibrillation (AF) and sinus (normal) rhythm. In other words AF versus not AF. This was the data set used for evaluation purposes with respect to the work presented in [1]. The AF/not AF class split was 60/60 records (see Table 1)

The GHS data set comprised the entire Guangzhou Heart Study data set of 1172 patients. This was the largest data set considered and therefore Feature Selection was applied, Stage 3 in the proposed process given in Figure 1. The individual patients held in the data set were categorised as being either Normal or Abnormal. The Normal/Abnormal class split was 878/294. Thus, unlike in the case of the GAF data set, the GHS data set was significantly imbalanced. Data augmentation was therefore also applied, Stage 4 in the proposed process given in Figure 1. It should also be noted that the GHS data set was comprised mostly of female patients, 773 (66%) compared to 399 males (34%). The age distribution was as follows: Normal class, 283 males and 595 females; Abnormal class, 116 males and 178 females. The age range of all the patients in the GHS data set was from 49 to 96, with a mean age 71.4 (a standard deviation of 6.260).

The LHCH data set was collected by the authors in collaboration with the Liverpool Heart and Chest Hospital. This data set focused on the recurrence of Atrial Fibrillation (AF) after catheter ablation which is estimated to be between 20% and 45% [10]. Accordingly, the data set comprised two classes: (i) patients who had AF and a catheter ablation where there was no recurrence, and (ii) patients who had AF and a catheter ablation where there was a recurrence. Details of all patients who had AF and a catheter ablation at the hospital, between June 2013 and December 2019, were recorded in a prospectively maintained data registry. For the LHCH data set patients were included if all their clinical and ECG data was available. ECG scanned images were only considered if they were taken within six months before the ablation. This meant that some patients had more than one ECG image associated with them. The minimum was one and the maximum was ten, but the average was two. In other words, the LHCH data set featured a "multiple learning classification" issue. Each image was stored using TIFF compression at a resolution of 300 dpi. In total, the LHCH data set comprised 909 patients and 1821 ECG images.

The LHCH data set also included information related to gender, age, body mass index (high times weight), and the presence of concomitant diseases. These features were all included because, according to Freming's study [36], these were risk factors related to AF recurrence. The following concomitant diseases were considered relevant: heart failure, hypertension, diabetes mellitus, hypercholesterolaemia, chronic kidney disease, thyroid dysfunction, and chronic obstructive pulmonary disease. These were selected because these had been identified in the study reported in [10]. The Left Atrial

(LA) size was also included, because this is considered to be an effective factor for predicting AF recurrence [27].

The final LHCH data set was composed of 610 male patients (67%) against 299 female patients (33%), distributed as follows: No recurrence class, 428 males and 211 females; Recurrence class, 182 males and 88 females. Thus, as in the case of the GHS data set, the LHCH was significantly imbalanced. Therefore, referring back to the proposed process (Figure 1), the application of Data Augmentation (Stage 4) was also applied to this data set.

Table 2: Descriptive Statistics for LHCH data set

Descriptive Statistics					
Risk factor	N	Minimum	Maximum	Mean	Std. Deviation
Age	909	19	84	60.5	10.711
Height	909	131	197	173.49	9.683
Weight	909	45.50	150	88.47	16.582
LA size	909	20	100	41.36	5.495

Of the total of 909 patients, 176 (19%) had never used alcohol while 733 (81%) were current drinkers of alcohol. Furthermore, 506 (56%) have never smoked, 66 (7%) were current smokers and 341 (37%) were ex-smokers. Despite the minimum age of the patients being 19 years, the mean age was 60.5 years, implying that there were more elderly patients in the sample than younger. Some further statistical detail concerning the LHCH data set is given in Table 2. From the foregoing, it is clear that the LHCH data set was the most sophisticated, in terms of additional features, of the three data sets considered.

5.2 Most Appropriate Feature Selection and Data Augmentation Techniques (Objective 1)

The proposed process, as described in Section 4, includes a feature selection stage (Stage 3) and a data augmentation stage (Stage 4). Neither was included in the original process presented in [1]. The first is used where overfitting occurs because a large number of motifs have been derived. Empirical evidence (not reported here) suggests that overfitting occurs when the number of motifs exceeds 3000. This was the case with respect to the GHS data set. The second was used where a significant class imbalance existed. This was the case with respect to both the GHS and LHCH data sets. Referring back to Sub-section 4.3 three feature selection techniques were suggested: (i) Principal Components Analysis (PCA), (ii) Singular Value Decomposition (SVD) and (iii) T-distributed Stochastic Neighbor Embedding (T-SNE). Referring back to Sub-section 4.4 three data augmentation techniques were suggested: (i) SMOTE, (ii) SVM-SMOTE and (iii) ADASYN. The operation of all these techniques were compared to determine the most appropriate.

Table 3 shows the results obtained using SVM classification and the three feature selection techniques considered when applied to the GHS data set (best results in bold font). SMOTE data augmentation was used in all three cases because further experiments (reported later in this sub-section) indicated that this produced the best results.

Inspection of the table indicates that best values were obtained using T-SNE, while the application of PCA resulted in overfitting. It was thus concluding that for feature selection T-SNE was the most appropriate choice.

Table 3: Evaluation results, using the GHS data set and SMOTE augmentation, to determine the most appropriate feature selection technique (best results in bold font)

DR Techniques	**Accuracy** *%*	**Precision** *%*	**Recall** *%*	**F1** *%*	**AUC** *%*
PCA	Overfitting				
SVD	71.52	65.10	93.20	76.58	71.58
TSNE	**81.50**	**84.00**	**78.15**	**80.87**	**81.54**

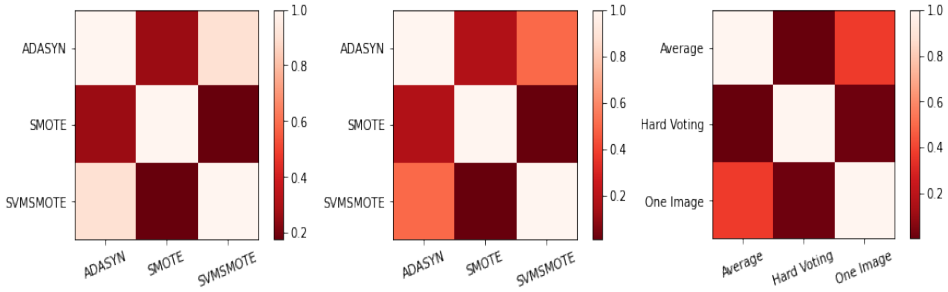
Table 4: Evaluation results, using the GHS and LHCH data sets and T-SNE feature selection, to determine the most appropriate data augmentation technique (best results in bold font)

Technique	**GHS data set**					**LHCH data set**				
	Accuracy *%*	**Precis.** *%*	**Recall** *%*	**F1** *%*	**AUC** *%*	**Accuracy** *%*	**Precis.** *%*	**Recall** *%*	**F1** *%*	**AUC** *%*
SMOTE	**81.50**	**84.00**	**78.15**	**80.87**	**81.54**	**66.82**	**74.89**	**52.17**	**60.62**	**66.83**
SVMSMOTE	78.49	79.86	84.86	82.21	77.17	62.76	57.49	97.15	72.15	62.67
ADASYN	79.24	82.86	75.11	78.68	79.40	64.51	62.47	80.66	70.15	63.78

Table 4 shows the results obtained using SVM classification and the three data augmentation techniques considered, when applied to the GHS and LHCH data sets (best results in bold font). T-SNE feature selection was used in all three cases because this had been shown to provide the best results (as reported in Table 3). For the LHCH data set an averaging technique was used for the multiple instance learning; later experiments, reported in 5.3, indicated that this produced the best results. From Table 4, it can be seen that using SMOTE produced best results. A subsequent Friedman Test indicated a statistically significant difference with respect to all the results obtained. Figures 2a and 2b show the outcomes obtained from consequent Nemenyi post-hoc tests for the two data sets (GHS and LHCH). From the figures, it can be seen that there was a statistically significant difference when using SMOTE compared with the other methods considered. It was thus concluding that for data augmentation SMOTE was the most appropriate choice.

5.3 Most Appropriate Conflict Resolution Technique (Objective 2)

As noted earlier in Sub-section 5.1, the LHCH data set, in many cases, features more than one ECG image for each patient resulting in Multiple Instance Classification requiring some form of conflict resolution should contradictory classifications result. In Sub-section 4.7 three conflict resolution techniques were suggested: (i) Averaging, (ii)



(a) Nemenyi post-hoc test for augmentation techniques using the GHS data set

(b) Nemenyi post-hoc test for augmentation techniques using the LHCH data set

(c) Nemenyi post-hoc test for conflict resolution using LHCH data set

Fig. 2: Nemenyi post-hoc test result

Table 5: Evaluation results, using the LHCH data sets, coupled with SMOTE data augmentation and T-SNE feature selection, to determine the most appropriate conflict resolution technique (best results in bold font)

DR Techniques	Accuracy %	Precision %	Recall %	F1 %	AUC %
Average	**66.82**	**74.89**	**52.17**	**60.62**	**66.83**
Voting	52.78	75.67	58.62	55.41	51.46
One Image	59.40	65.57	40.43	49.63	59.45

Voting and (iii) avoiding the problem by using only the most recent ECG image. The second evaluation objective, Objective 2, was to identify which of these techniques produced the most effective classification results. The results are presented in Table 5 (best results in bold font). From the table, it can be seen that the averaging produced the best results. A Friedman Test indicated a statistically significant difference in the results. The outcome of a consequent Nemenyi post-hoc test is presented in Figure 2c. From the figure, it can be seen that there was a clear statistically significant difference in operation when using averaging.

5.4 Operation Using Additional Features (Objective 3)

The previous sub-section described the experiments conducted to determine the best techniques to be used for: (i) feature selection, (ii) data augmentation and (iii) conflict resolution (in the event of multiple instance classification). Best results were obtained using: T-SNE feature selection, SMOTE data augmentation and averaging (where applicable). In this section, the evaluation results obtained from further experiments, conducted using additional features, are discussed. Combinations of: motifs, discords, and clinical data such as age and gender. The aim was to determine whether any advantage would be gained by adding additional features from related sources. Similar experiments were conducted in [2]; but when using 1D time series extracted from ECG traces. The reported results indicated that adding additional features improved the effectiveness

Table 6: Evaluation results, coupled with (where required) SMOTE data augmentation, T-SNE feature selection and averaging conflict resolution, to determine the most appropriate data combination (best results in bold font)

Evaluation of Proposed Approach						
Data set	**HFVR**	**Accuracy %**	**Precision %**	**Recall %**	**F1 %**	**AUC %**
GAF Data set	Motifs only	84.6	84.00	86.21	83.71	85.61
	Discords only	45.00	45.00	43.00	43.98	46.50
	Clinical only	56.25	55.50	54.33	53.85	54.67
	Motifs+ Discords	77.50	73.48	85.14	76.78	77.99
	Motifs+Clinical	**86.25**	**85.83**	**86.50**	**84.88**	**85.87**
	Discords +Clinical	48.75	51.70	65.67	55.23	51.42
	Motifs+Discords +Clinical	78.75	74.83	82.57	77.00	79.62
GHS Data set	Motifs only	81.50	84.00	78.15	80.87	81.54
	Discords only	52.43	52.36	60.69	55.84	52.77
	Clinical only	64.08	64.85	61.10	62.80	63.99
	Motifs+ Discords	72.76	74.50	69.54	71.80	72.74
	Motifs+Clinical	**84.09**	**87.76**	**79.20**	**83.17**	**84.05**
	Discords +Clinical	59.87	60.27	59.65	59.56	59.95
	Motifs+Discords +Clinical	82.58	87.89	75.74	81.24	82.56
LHCH Data set	Motifs only	66.82	74.89	52.17	60.62	66.83
	Discords only	46.63	47.80	58.61	51.79	47.45
	Clinical only	82.32	78.88	88.22	83.16	82.43
	Motifs+ Discords	66.89	66.31	76.47	69.44	67.60
	Motifs+Clinical	**84.59**	**79.84**	**92.29**	**85.57**	**84.56**
	Discords +Clinical	76.61	81.16	69.87	74.79	76.81
	Motifs+Discords +Clinical	81.38	86.14	74.89	80.01	81.43

of the CVD classification. Further motivation was provided from work, such as that reported in [17, 28], which suggested that age, gender, smoking status, and so on, were all risk factors to be considered when classifying ECG data. Experiments were conducted using all three data sets, the GAF, GHS and LHCH. In each case seven different data combinations were considered: (i) motifs only (M), (ii) discords only (S), (iii) clinical data only (C), (iv) motifs and discords (M+S), (v) motifs and clinical data (M+C), (vi) discords and clinical data (S+C) and (vii) motifs, discords and clinical data (M+S+C). In each case, were appropriate, T-SNE feature selection, SMOTE data augmentation, and averaging conflict resolution were used.

The results are presented in Table 6. From the table, it can be observed that in all three cases, the combination of motifs and clinical data produced the best results. The worst results were obtained using discords. Indeed, from the results obtained, it can be argued that the inclusion of discords had a negative effect as evidenced when discords were added to the motif and clinical data combination. It should also be noted here, with respect to the results presented in Table 6, that we re-ran the experiments for the GAF data set to calculate AUC, Unlike the case of the evaluation reported in [1] repeated Ten-cross validation was used, rather than single Ten-cross validation. Consequently, the results presented in Table 6 are not identical to the ones presented previously in

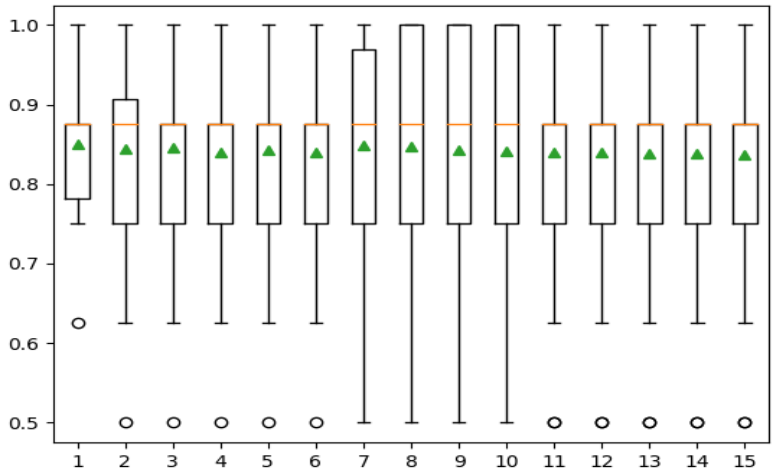


Fig. 3: Box and Whisker plots of the accuracy vs fifteen repeats for Ten-fold-cross-validation

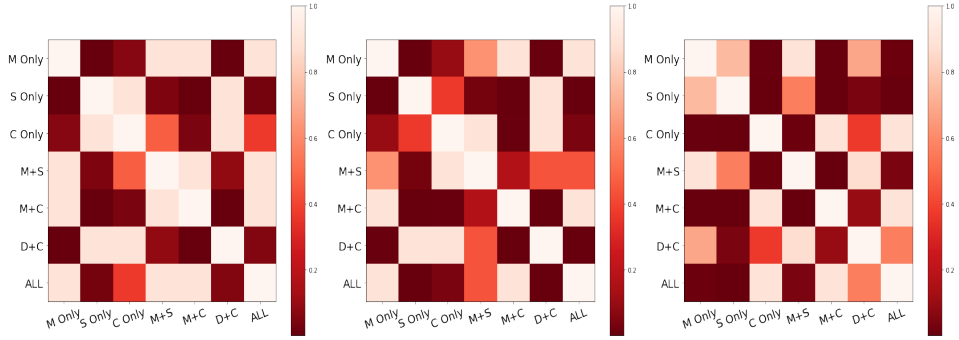
[1]. The rational for using repeated Ten-fold cross-validation was that a more reliable estimate of model performance would be obtained. Note that we calculated the mean, and the standard error of the accuracy on each iteration to minimize the standard error and stabilize the mean of estimated performance.

Figure 3, shows a sequence of Box and Whisker plots for the recorded accuracy for the motifs-only model. The Y-axis gives the accuracy and the X-axis the number of cross validation repeats. From the figure, it can see that the mean seems to be around a value of 84.6 and that the standard error decreased with the increase in the number of repeats and stabilized with a value around 0.010.

A Friedman test was also applied with respect to the results obtained using each data set. The Friedman test demonstrated that there was a statistically significant difference in performance in all three cases. The results of the consequent Nemenyi post-hoc tests are presented in Figure 4.

Figure 4a presents the Nemenyi post-hoc test results using the GAF data set. From the figure, it can be seen that there is a statistically significant difference when using motifs combined with clinical data compared to many of the other data combinations considered. The exceptions were the motifs and discords; and the motifs, discords and clinical data combinations.

Figure 4b shows the Nemenyi post-hoc test results using the GHS data set. The best overall results were obtained using the GHS data set, and particularly when motifs were combined with clinical data (an AUC of 84.05%). From the figure it can be seen that the results obtained when using motifs combined with clinical data were statistically different from most of the other combinations considered; with the exception of motifs used on their own, and motifs coupled with discords and clinical data.



(a) GAF data set (b) GHS data set (c) LHCH data set

Fig. 4: Nemenyi post-hoc test when using different data combinations

Figure 4c presents the Nemenyi post-hoc test results using the LHCH data set. From the figure, it can be seen that there is also a statistically significant difference when using motifs combined with clinical data compared to many of the other data combinations considered. The exceptions were the clinical only, and the motifs coupled with discords and clinical data.

5.5 Comparison of 1D and 2D motifs Discovery Approaches (Objective 4)

In [1] a comparison between 1D (time series) and 2D (image) approaches for the GAF data set, using motifs on their own, discords on their own, and motifs and discords combined. These results are presented in the top part of Table 7. The comparison was made to investigate the hypothesis that using 2D motifs extracted from untransformed ECG images would produce a better classification than that obtained using features selected from 1D transformed waveform representations of ECG data. For the work presented here this experiment was repeated using the GHS and LHCH data sets. To obtain the ID results the scanned GHS and LHCH images were transformed into a time series format using a recent algorithm for achieving this [12]. Following the transformation, the 1D motif approach proposed in [3] was applied. The results are presented in the lower part of Table 7. In the table best results in each case are highlighted in bold font. From the table, it can be seen that the outcomes from the experiments using GHS and LHCH data stes corroborated the results reported in [1]. That the 2D formatted data produced better results than the 1D format. Interestingly, from Table, the 1D waveform approach that using a combination of motifs as features, works well in comparison to other 1D waveform approaches. However, from Table 7, best results for the GAF and GHS data sets were obtained using 2D motifs only; while for the LHCH data set best results were obtained using 2D motifs and discords combined.

6 Conclusion

In this paper, the approach to ECG scanned image classification using 2D motifs reported in [1] has been extended and re-analysed using additional data sets and evaluation aspects. It was assumed that the “traditional” approach to ECG classification

Table 7: Comparison of 1D and 2D motifs Discovery Approaches

		2D Approach					**1D Approach**				
Data set	**HFVR**	**Acc.** *%*	**Pre.** *%*	**Rec.** *%*	**F1** *%*	**AUC** *%*	**Acc.** *%*	**Pre.** *%*	**Rec.** *%*	**F1** *%*	**AUC** *%*
GAF	Motifs only	**85.00**	**84.00**	**86.21**	**83.71**	**85.61**	68.48	70.00	68.49	69.88	69.28
Data set	Discords only	45.00	45.00	43.00	43.98	46.50	67.59	76.00	66.59	71.24	68.23
	Motifs + Discords	77.50	73.48	85.14	76.78	77.99	**72.35**	**78.74**	**72.50**	**75.49**	**71.92**
GHS	Motifs only	**81.50**	**84.00**	**78.15**	**80.87**	**81.54**	72.119	77.96	70.01	73.22	73.00
Data set	Discords only	52.43	52.36	60.69	55.84	52.77	69.59	74.05	68.81	70.53	70.36
	Motifs + Discords	72.76	74.50	69.54	71.80	72.74	**76.16**	**83.56**	**73.28**	**77.66**	**77.17**
LHCH	Motifs only	66.82	74.89	52.17	60.62	66.83	64.69	72.60	62.96	67.08	65.41
Data set	Discords only	46.63	47.80	58.61	51.79	47.45	64.14	72.32	62.20	66.36	65.23
	Motifs + Discords	**66.89**	**66.31**	**76.47**	**69.44**	**67.60**	**66.76**	**73.77**	**64.96**	**68.70**	**66.81**

using waveform transformation and limited features resulted in information loss due to the approximations used, and that a better classification could be obtained if the classification model was built using the original image without any transformations. To investigate this, three data sets were tested using the 2D motifs approached. The potential of including other clinical features such as age and gender were also investigated and it was found that this provided better results. The reported evaluation demonstrated that the best results were obtained when 2D motifs were extracted from an entire image compared with when the image was transformed into a 1D waveform format and 1D motifs used as features. The best accuracy of 85% was obtained using the proposed approach, and 86.25% when adding additional clinical features, in comparison with the best accuracy of 68.48% using the 1D waveform format. For future work, the authors intend to investigate: (i) improving the performance of the 2D motif extraction from scanned images process, (ii) the effect of combining 2D motifs and discords with features from other formats such as Echo data and patient data, and (iii) the application of the proposed approach to alternative CVD application domains that feature multi-class classification.

References

1. Aldosari., H., Coenen., F., Lip., G.Y.H., Zheng., Y.: Two-dimensional motif extraction from images: A study using an electrocardiogram. In: Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR,. pp. 19–28. INSTICC, SciTePress (2022)
2. Aldosari, H., Coenen, F., Lip, G.Y., Zheng, Y.: Addressing the challenge of data heterogeneity using a homogeneous feature vector representation: A study using time series and cardiovascular disease classification. In: International Conference on Innovative Techniques and Applications of Artificial Intelligence. pp. 254–266. Springer (2021)
3. Aldosari, H., Coenen, F., Lip, G.Y., Zheng, Y.: Motif based feature vectors: towards a homogeneous data representation for cardiovascular diseases classification. In: International Conference on Big Data Analytics and Knowledge Discovery. pp. 235–241. Springer (2021)
4. Apostolico, A., Parida, L., Rombo, S.E.: Motif patterns in 2d. Theoretical Computer Science **390**(1), 40–55 (2008)

5. Badilini, F., Erdem, T., Zareba, W., Moss, A.J.: Ecgscan: a method for conversion of paper electrocardiographic printouts to digital electrocardiographic files. Journal of electrocardiology **38**(4), 310–318 (2005)
6. Baydoun, M., Safatly, L., Abou Hassan, O.K., Ghaziri, H., El Hajj, A., Isma'eel, H.: High precision digitization of paper-based ecg records: a step toward machine learning. IEEE journal of translational engineering in health and medicine **7**, 1–8 (2019)
7. Bosch, A., Munoz, X., Marti, R.: Which is the best way to organize/classify images by content? Image and vision computing **25**(6), 778–791 (2007)
8. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority oversampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)
9. Deng, H., Guo, P., Zheng, M., Huang, J., Xue, Y., Zhan, X., Wang, F., Liu, Y., Fang, X., Liao, H., et al.: Epidemiological characteristics of atrial fibrillation in southern china: results from the guangzhou heart study. Scientific reports **8**(1), 1–10 (2018)
10. Dretzke, J., Chuchu, N., Agarwal, R., Herd, C., Chua, W., Fabritz, L., Bayliss, S., Kotecha, D., Deeks, J.J., Kirchhof, P., et al.: Predicting recurrent atrial fibrillation after catheter ablation: a systematic review of prognostic models. EP Europace **22**(5), 748–760 (2020)
11. Ebrahimi, Z., Loni, M., Daneshtalab, M., Gharehbaghi, A.: A review on deep learning methods for ECG arrhythmia classification. Expert Systems with Applications: X **7**, 100033 (2020)
12. Fortune, J., Coppa, N., Haq, K.T., Patel, H., Tereshchenko, L.G.: Digitizing ECG image: new fully automated method and open-source software code. medRxiv (2021)
13. Furfaro, A., Groccia, M.C., Rombo, S.E.: 2d motif basis applied to the classification of digital images. The Computer Journal **60**(7), 1096–1109 (2017)
14. Gupta, V., Mittal, M., Mittal, V., Saxena, N.K.: A critical review of feature extraction techniques for ECG signal analysis. Journal of The Institution of Engineers (India): Series B pp. 1–12 (2021)
15. Houssein, E.H., Kilany, M., Hassanien, A.E.: ECG signals classification: a review. International Journal of Intelligent Engineering Informatics **5**(4), 376–396 (2017)
16. Jayaraman, S., Swamy, P., Damodaran, V., Venkatesh, N.: A novel technique for ECG morphology interpretation and arrhythmia detection based on time series signal extracted from scanned ECG record. Advances in Electrocardiograms-Methods and Analysis pp. 127–140 (2012)
17. Joseph, P., Kutty, V.R., Mohan, V., Kumar, R., Mony, P., Vijayakumar, K., Islam, S., Iqbal, R., Kazmi, K., Rahman, O., et al.: Cardiovascular disease, mortality, and their associations with modifiable risk factors in a multi-national south asia cohort: a pure substudy. European Heart Journal **43**(30), 2831–2840 (2022)
18. Kar, A., Das, L.: A technical review on statistical feature extraction of ecg signal. In: IJCA Special Issue on 2nd National Conference-Computing, Communication and Sensor Network, CCSN. pp. 35–40 (2011)
19. Khleaf, H.K., Ghazali, K.H., Abdalla, A.N.: Features extraction technique for ECG recording paper. In: Proceeding of the International Conference on Artificial Intelligence in Computer Science and ICT (2013)
20. Liu, B., Li, J., Chen, C., Tan, W., Chen, Q., Zhou, M.: Efficient motif discovery for large-scale time series in healthcare. IEEE Transactions on Industrial Informatics **11**(3), 583–590 (2015)
21. Liu, X., Wang, H., Li, Z., Qin, L.: Deep learning in ECG diagnosis: A review. Knowledge-Based Systems **227**, 107187 (2021)
22. Loresco, P.J.M., Africa, A.D.: ECG print-out features extraction using spatial-oriented image processing techniques. Journal of Telecommunication, Electronic and Computer Engineering (JTEC) **10**(1-5), 15–20 (2018)

23. Maletzke, A.G., Lee, H.D., Batista, G.E., Rezende, S.O., Machado, R.B., Voltolini, R.F., Maciel, J.N., Silva, F.: Time series classification using motifs and characteristics extraction: a case study on ECG databases. In: Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support. pp. 322–329. Atlantis Press (2013)
24. Mensah, G.A., Roth, G.A., Fuster, V.: The global burden of cardiovascular diseases and risk factors: 2020 and beyond. Journal of the American College of Cardiology **74**(20), 2529–2532 (2019)
25. Mir, H.Y., Singh, O.: ECG denoising and feature extraction techniques–a review. Journal of medical engineering & technology **45**(8), 672–684 (2021)
26. Nadakinamani, R.G., Reyana, A., Kautish, S., Vibith, A.S., Gupta, Y., Abdelwahab, S.F., Mohamed, A.W.: Clinical data analysis for prediction of cardiovascular disease using machine learning techniques. Computational Intelligence and Neuroscience, Soeial Issue on Artificial Intelligence and Machine Learning-Driven Decision-Making (2022)
27. Njoku, A., Kannabhiran, M., Arora, R., Reddy, P., Gopinathannair, R., Lakkireddy, D., Dominic, P.: Left atrial volume predicts atrial fibrillation recurrence after radiofrequency ablation: a meta-analysis. Ep Europace **20**(1), 33–42 (2018)
28. Peters, S.A., Wang, X., Lam, T.H., Kim, H.C., Ho, S., Ninomiya, T., Knuiman, M., Vaartjes, I., Bots, M.L., Woodward, M.: Clustering of risk factors and the risk of incident cardiovascular disease in asian and caucasian populations: results from the asia pacific cohort studies collaboration. BMJ open **8**(3), e019335 (2018)
29. Ravichandran, L., Harless, C., Shah, A.J., Wick, C.A., Mcclellan, J.H., Tridandapani, S.: Novel tool for complete digitization of paper electrocardiography data. IEEE journal of translational engineering in health and medicine **1**, 1800107–1800107 (2013)
30. Roth, G.A., Mensah, G.A., et al., C.O.J.: Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the gbd 2019 study. Journal of The American College of Cariology (25), 2982–3201 (2020)
31. Sahoo, S., M.Dash, S.Behera, S.Sabut: Machine learning approach to detect cardiac arrhythmias in ECG signals: A survey. Innovation and Research in BioMedical Engineering **41**(4), 185–194 (2020)
32. Seena, V., Yomas, J.: A review on feature extraction and denoising of ECG signal using wavelet transform. In: 2014 2nd international conference on devices, circuits and systems (ICDCS). pp. 1–6. IEEE (2014)
33. Smíšek, R.: ECG signal classification based on svm. Biomedical Engineering (1), 365–369 (2016)
34. Thanapatay, D., Suwansaroj, C., Thanawattano, C.: Ecg beat classification method for ecg printout with principle components analysis and support vector machines. In: 2010 International Conference on Electronics and Information Engineering. vol. 1, pp. V1–72. IEEE (2010)
35. Torkamani, S., Lohweg, V.: Survey on time series motif discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **7**(2), e1199 (2017)
36. Truong, C.D., Nguyen, B.T., Van Cong Tran, T.: Prediction of risk factors for recurrence of atrial fibrillation in patients with arterial hypertension. The International Journal of Cardiovascular Imaging **37**(12), 3413–3421 (2021)
37. Velliangiri, S., Alagumuthukrishnan, S., et al.: A review of dimensionality reduction techniques for efficient computation. Procedia Computer Science **165**, 104–111 (2019)
38. Wankhedkar, R., Jain, S.K.: Motif discovery and anomaly detection in an ECG using matrix profile. In: Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2019, Volume 1. pp. 88–95. Springer (2021)