Addressing the Challenge of Data Heterogeneity using a Homogeneous Feature Vector Representation: a Study using Time Series and Cardiovascular Disease Classification

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

[1] Department of Computer Science,
[2] Liverpool Centre for Cardiovascular Science,
[3] Department of Eye and Vision Science, University of Liverpool, Liverpool, UK.
{H.A.Aldosari,Coenen,Gregory.Lip,yalin.zheng}@liverpool.ac.uk

Abstract. An investigation into the use of a unifying Homogeneous Feature Vector Representation (HFVR), to address the challenge of applying machine learning and/or deep learning to heterogeneous data, is presented. To act as a focus, Atrial Fibrillation classification is considered which features both tabular and Electrocardiogram (ECG) time series data. The challenge of constructing HFVRs is the process for selecting features. A mechanism where by this can be achieved, in terms of motifs and discords, with respect to ECG time series data is presented. The presented evaluation demonstrates that more effective AF classification can be achieved using the idea of HFVR than would otherwise be achieved.

Keywords: Unifying Homogeneous Feature Vector Representations · Time Series Feature Extraction and Analysis · Atrial Fibrillation Classification.

1 Introduction

The sophistication of the global technical infrastructure, and the consequent data acquisition capabilities, are rapidly growing and producing large amounts of data. There is a corresponding interest in strategies and techniques for the automated abstraction of knowledge and analysis of this data. Strategies and techniques that typically employ Machine Learning and Deep Learning (ML/DL).

However, these techniques and strategies still face many challenges. One such challenge, and that of relevance with respect to the work presented in this paper, is that of mixed data formats (types), also known as heterogeneous data. A trivial illustration of this is the distinction between numeric data (for example age) and categorical data (for example gender). A more comprehensive example is the distinction between video and free text data. A range of solutions have been proposed, which can be categorised as follows:

1. **Direct Conversion**. Given data sources in several formats convert all the data into one of the formats used, and then apply the ML/DL.

2. **Independent ML/DL Application**. Apply the ML/DL to each format independently and then combine the results.
3. **Unification**. Create a unifying representation that sits over the input formats and then apply the ML/DL to this unified format.

The distinct between the above is illustrated in Figure 1.

The conversion solution is the most straightforward solution. The solution is based on the observation that numerical values can be converted into categorical values (discretisation) [25], and that categorical values can be converted into numerical values (normalisation) [8, 7]. However, the conversion solution assume that given data in two formats it is possible to convert one into the other; this is generally not possible without introducing significant simplification and/or approximation, which in turn may affect the consequent ML/DL.

The independent application of ML/DL solution considers each data format independently; ML/DL is applied with respect to each format, and the "local" results combined to give a "final" global result. For example, given a prediction model generation problem, which has as input data in several formats, build prediction models with respect to each format; and then, given a previously unseen record, apply the models and combine the local predictions to give a global prediction using (say) voting as used in ensemble classification systems [10, 14]. This idea has also been incorporated into the concept of "multi-input networks" [23, 24] where each format is associated with a different branch of the network. The branches are then brought together at the end so that a single final result will be produced. However, the idea of using multiple ML/DL applications ignores the existence of any relationships that might exit across the data sources, which in turn may have an adverse effect on the quality of the ML/DL.

The idea underpinning the unification solution to the heterogeneous source data problem is that of creating a unifying representation that sits over the heterogeneous data sources. This then serves to capture the relationships that may exist across the data; not the case with the independent ML/DL solution. This idea was promoted in [1] where a Homogeneous Feature Vector Representation (HFVR) was proposed in the context of classification model generation. The intuition was that a feature vector representation is compatible with a wide range of classification model generators. However the main challenge of the HFVR approach is how best to identify the features (attributes) to be included in the HFVR; these should be features that are good discriminators of class. In [1] a Cardiovascular Disease (CVD) scenario was considered that featured time series data, numeric data and categorical data. Motifs were extracted from Electrocardiograms (ECGs) and used to create an HFVR. Reasonable results were obtained. A criticism of the work in [1] is that only one type of time series feature was used, motifs (frequently occurring time series sub-sequences). It is argued here that a more sophisticated HFVR would have been produced, and hence better classification results produced, if more than one time series feature type were considered.

In this paper the idea of HFVR is explored further, building on the work presented in [1] also using a CVD scenario, more specifically Atrial Fibrillation

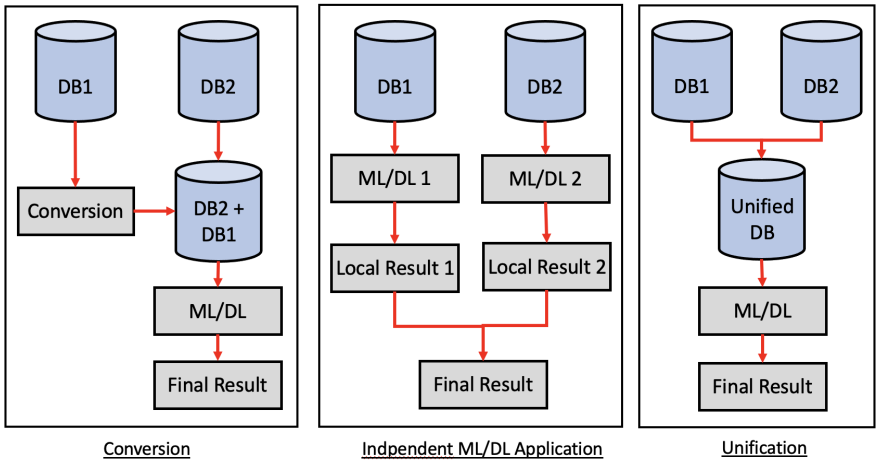

Fig. 1. Distinctions between the potential solutions to the heterogeneous source data ML/DL challenge.

(AF) classification, as the application focus for the work. The research hypothesis considered is that for the concept of HFVR to operate effectively much depends on the nature of the features that are included. With respect to some data formats (such as free text, graphs, image, video and time series), this is a non trivial task; with others this is reasonably straight forward (for example numeric or categorical tabular data). The work presented in this paper thus focuses on feature extraction from time series and combining this with numeric and categorical data to form an HFVR. To this end, an HFVR generation approach is presented where by both time series motifs and discords (anomalous time series sub-sequences), coupled with numeric and categorical data, are used to form an HFVR. The reported evaluation indicates that the above hypothesis is correct, and that a more effective prediction system can be built using a more sophisticated set of HFVR features than in earlier work; at least in the context of AF classification.

The remainder of this paper is structured as follows. A review of existing work relevant to this paper is presented, in Section 2. The AF application domain used as a focus for the work presented in this paper is presented in Section 3 and a number of relevant definitions in Section 4. The proposed process for generating an HFVR in the context of AF classification is presented in Section 5, and the associated evaluation in Section 6. The paper is concluded in Section 7 with a summary of the main findings and some suggestions for future work.

2 Previous Work

The domain of interest with respect to this paper is Cardiovascular Disease (CVD) analysis. CVD analysis typically relies on Electrocardiogram (ECG) data coupled with tabular data. ECGs are essentially time series. To build an HFVR

that combines time series data with other data formats a feature extraction approach needs to be applied to the time series data.

Feature extraction is well established process in ML/DL used to: (i) reduce the complexity of ML/DL problems and (ii) enhance the effectiveness of the ML/DL models produced [13]. Feature extraction has been applied to time series with respect to many application domains, for example in [4] a time series feature extraction process in the context of a brain-computer interface application was described; and in [2] time series feature extraction was used in the context of signals captured from "rotating machines" (gear boxes). Time series feature extraction has also, as might be expected, been applied to ECG time series data [11]. In this context feature extraction has typically been applied with respect to the global characteristics of ECG time series [18, 26]. The global characteristics extracted from ECG data are typically time intervals between key events, and the amplitudes associated with key events with respect to ECG heart beat cycles. A range of techniques have been proposed to achieve this, such as wavelet transform (WT) [5]. There are more than ten global features that can be extracted, some approaches only extract two [21], some four [5], others eight [15]. Intuitively, the more features extracted the better the classification accuracy achieved. However, to reduce the computational overhead associated with ML/DL model generation, while maintaining accuracy, feature selection can be applied subsequent to the feature extraction process [21].

An alternative to extracting global characteristics as features from time series is to extract discriminative time series sub-sequences. One popular form of discriminative time series sub-sequence is the *motif*, defined as a frequently occurring sub-sequence which, it is assumed, is therefore representative of the time series [19]. The advantage offered, compared to the use of global characteristics as features, is that they are easy to identify. The basic idea is to extract all-time series sub-sequences of length w from a "parent" time series, and then select the most frequently occurring. To count the frequency of occurrence we cannot require sub-sequences to match exactly for them to be said to co-occur as very few exact matches occur in practice. Instead a similarity threshold is required applied to a similarity measure of some form. A range of techniques have been proposed to extract motifs from ECG data. In [18] both motifs and global characteristics were extracted from ECG data. Motifs were used with respect to the work presented in [1], which underpins the work presented in this paper.

Motifs are not the only time series sub-sequences that can be used as features. An alternative, and that of significance with respect to this paper, is discords [12]. A discord is a unique sub-sequence within a time series which, it is assumed, is therefore representative of the time series; the counter-argument to the argument for motifs. The work presented in this paper investigates the use of motifs and discords, coupled with numeric and categorical data, to create an HFVR to which ML/DL can be applied.

A variety of mechanisms have been proposed whereby discords and motifs can be discovered. A popular method is the Matrix Profile (MP) technique first proposed in [28]. A MP is a data structure which features two vectors for each

time series; a Distance Profile (DP) and Profile Index (PI). The idea is to first extract all sub-sequences from a time series using a sliding window of length w. Then to determine the pairwise distances between the sub-sequences and store these in a matrix. However, the matrix will include redundant information. Only the distances between each sub-sequence and its nearest neighbour will be required (the minimum distance). These distances are therefore stored in a DP and the index of the neighbouring time series in a PI. These two vectors then facilitate the extraction of motifs and discords, of a given length w, in a manner that both avoids approximation and is much more efficient than previously proposed approaches. The Matrix Profile MP idea was used, in the context of ECG time series analysis, in [27]. A number of alternative examples can be found in the literature concerning techniques to extract motifs and discords from ECG data, such as the Multivariate Maximal Time Series technique used in [22]. However, the matrix profile idea is used with respect to the work presented in this paper.

A number of algorithms have been proposed to compute MPs, examples include: the Scalable Time Series Anytime Matrix Profile (STAMP) algorithm [28] and the Scalable Column Independent Matrix Profile (SCRIMP) algorithm [30]. For the work presented in this paper the Correct Matrix Profile (CMP) mechanism, described in [6], was adopted; an extension of the MP technique. The idea here, once the MP has been computed by one of the above algorithms, is to add an Annotation Vector (AV) which contains ranking values of between 0 and 1 effectively changing the shape of the MP so that the motifs and discords are ordered and hence the top-K can be selected.

3 Application Domain

The application focus for the work presented in this paper, as already noted, is CVDs. CVDs are diseases of the heart and blood vessels. According to the World Health Organisation (WHO), 17.9 million people die each year from CVDs, an estimated 31% of all deaths worldwide. The domain of CVD analysis is extensive, therefore the focus of the work presented in this paper is Atrial Fibrillation (AF). AF is a heart condition where the *atria*, the upper two chambers of the heart, contracts in an abnormal manner. Because of the irregular beating of the heart, blood does not flow in a normal manner, and the electrical impulses that control the timing of the heart are disturbed [16]. This can be identified from range of tests including ECG analysis which is widely considered to be the most reliable test for diagnosing AF [3, 29]. An ECG indicates the electrical activity of the heart in terms of a summation wave that can be visualised and hence interpreted.

4 Formalism

The following definitions are used with respect to the remainder of this paper. Note that some of the definitions are specific to the CVD AF application focus considered.

Homogeneous Feature Vector Representation (HFVR): A data set $H = \{V_1, V_2, \ldots\}$, where each vector $V_i = \{v_1, v_2, \ldots\}$ comprises a set of values that correspond to selected features extracted from a heterogeneous data set. When used for model generation, each vector will include an associated class label c drawn from a set of classes C ($V_i = \{v_1, v_2, \ldots, c\}$).

Patient Record: A set of records $\mathbf{D} = \{R_1, R_2, \ldots\}$ where each record R_i comprises information about the patient. In the case of the CVD AF application domain this will include ECG data as well as more general patient data.

Time series: A collection of time series $\mathbf{T} = \{T_1, T_2, \ldots\}$, associated with a patient, representing ECG data. Each time series T_j is comprised of a sequence of data values $[t_1, t_2, \ldots, t_n]$.

Discords: A discord s is a time series sub-sequence t_j. S is a set of discords extracted from a collection of time series $\mathbf{T}$, $S = \{s_1, s_2, ...\}$. Not all discords in S will be good discriminators of class, so we prune S to give S' and then S'' (see below).

Motifs : A motif m is a time series sub-sequence t_j. M is a set of motifs extracted from collection of time series $\mathbf{T}$, $M = \{m_1, m_2, ...\}$. Again, not all the motifs in M will be good discriminators of class, so we prune M to give M' and then M''.

For the evaluation presented later in this paper a binary classification scenario is considered, $C = \{true, false\}$, "has AF" or "does not have AF". To generate the desired HVFR the input time series ECG data was first separated from the rest of the input to form a ECG time series data set $\mathbf{T} = \{T_1, T_2, \ldots\}$. Motifs and discords were then extracted from $\mathbf{T}$ and then combined with other data to form the desired HFVR. Each vector in the HFVR, representing a patient, thus comprised $\{S, M, A, c\}$ where A is the additional numeric and categorical patient data, and c is a class label taken from a set of classes C ($c \in C$). For the evaluation presented later in this paper $C = \{AF, \neg AF\}$; hence a binary classification. How the HFVR was generated, with reference to the CVD AF classification application, is discussed in the following section.

5 Homogeneous Feature Vector Representation Generation

This section presents the proposed HFVR generation mechanism. In the context of the CVD AF application the proposed process is as shown in Figure 2. In more detail, the CVD AF classification application included data from multiple data sources; data from electrocardiograms, blood tests, x-rays and clinical data. For each data set the features to be included in the HFVR needed to be extracted. In this paper, two data sets were used, ECG time series from which motifs and discords could be extracted and clinical data.

The adopted approach for motif and discords discovery (extraction and selection), as noted earlier (Section 2), was the CMP technique described in [6] using the Guided Motif Search (GMS) algorithm presented in [6]. Using the CMP

technique, motifs and discord sub-sequences are ranked. The top three motifs and discords were then selected. The proposed process to extract the features and create the desired HFVR was as follows:

1. Divide the input time series data $\mathbf{T} = \{T_1, T_2, \ldots\}$ into two D_1 and D_2, where D_1 corresponds to class c_1 and D_2 corresponds to class c_2 (note that we are assuming a binary classification here).
2. Extract the top three motifs for each $T_i \in D_1$ and each $T_i \in D_2$ to give M_1 and M_2 respectively.
3. Extract the top three discords for each $T_i \in D_1$ and each $T_i \in D_2$to give S_1 and S_2 respectively.
4. Compare the set of motifs within M_1(M_2) with each other, and select those that occur at least k times to give M_1' (M_2').
5. Compare the set of discords within S_1 (S_2) with each other, and retain those that occur at least k times to give S_1' (S_2').
6. There may be motifs that occur in both M_1' and M_2' that are associated with both class c_1 and c_2, and are therefore not good discriminators of class, prune these to give M''.
7. There may be discords that occur in both S_1' and S_2' that are associated with both class c_1 and c_2 and are therefore not good discriminators of class, prune these to give S''.
8. Use the content of M'' and S'', coupled with data attributes from A (the set of known numeric and categorical patient attributes) to create an HFVR.

For the evaluation presented in Section 6 below, to compare discords (motifs), Euclidean distance similarity was used with a similarity threshold σ. If the distance between two discords (motifs) was less than σ they were deemed to match.

6 Evaluation

The evaluation of the proposed approach to HFVR generation, in the context of CVD AF classification, was conducted using the China Physiological Signal Challenge 2018 (CPSC2018) data set[1]. Some detail concerning this data set are presented in Sub-section 6.1. The objectives of the evaluation were:

1. To determine whether the idea of a unifying HFVR did indeed provide real benefits in the context of the CVD AF classification application used as a focus with respect to this paper. In other words, that the use of a combined set of features, facilitated by the HFVR approach, did indeed provide for a more effective classification.
2. To provide evidence that the proposed approach to HFVR generation with respect to time series, using the CMP technique coupled with GMS, produced the desired result.

[1] http://2019.icbeb.org/Challenge.html

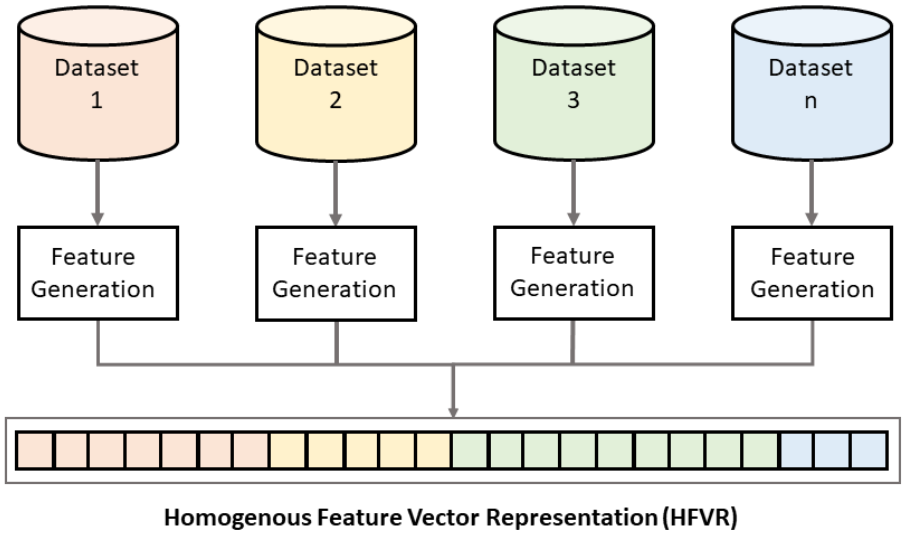

Fig. 2. Schematic of HFVR Generation Process in the context of the CVD AF classification exemplar application domain.

For the evaluation $\sigma = 0.15$ and $k = 70$ were used for discovering discords, and $\sigma = 0.15$ and $k = 90$ for discovering motifs. These were used because earlier experiments, not reported here for reasons of space, had demonstrated that these were the most appropriate values. The Support Vector Machine (SVM) classification model was used, coupled with the grid search technique to identify best parameters. A range of HFVRs were generated using combinations of features: motifs and discords generated from ECG time series and patient age and gender. The later were chosen because they were exemplars of numeric and categorical data features, and because they had been shown to have an impact on AF (see [20] and [9] respectively). Ten-fold Cross Validation (TCV) was used throughout. The metrics recorded were accuracy, precision, recall and F1; of which F1, the harmonic mean of precision and recall, is a good summarising measure. Experiments were conducted using the following four groupings:

1. **Group 1 - Gender and Age**: (i) gender only, (ii) age only, (iii) age + gender
2. **Group 2 - Motifs**: (i) motifs only, (ii) motifs + age, (iii) motifs + gender, and (iv) motifs + age + gender
3. **Group 3 - Discords**: (i) discords only, (ii) discords + age, (iii) discords + gender, and (iv) discords + age + gender.
4. **Group 4 - Motifs and Discords**: (i) motifs + discords, (ii) motifs + discords + age, (iii) motifs + discords + gender, and (iv) motifs + discords + age + gender.

Note that an HFVR comprised of one feature type was equivalent to using a standard feature vector representation. The results are presented and discussed in Sub-section 6.2.

6.1 The CPSC2018 Data Set

The CPSC2018 data set was curated for the cardiovascular disease detection competition held during the 7th International Conference on Biomedical Engineering and Biotechnology [17]. The CPSC2018 data comprises 6,877 digitised ECG records (3178 female and 3699 male), ranging from 6 to 60 seconds in duration and sampled at 500Hz. Each ECG recording has two files: (i) a binary file for the ECG signal data and (ii) a text file (header format) describing the recording and patient attributes, including age, gender and the diagnosis label (the class attribute c). The class labels were drawn from a set of eight arrhythmia types ($|C| = 8$). However, for the evaluation presented here only AF and Normal rhythm ($\neg$AF) were used. The ECG data was also rationalised so that all records were of fixed length, 5000 points. Each ECG time series was considered in isolation. For the evaluation presented here a total, 600 records were used (300 for each class).

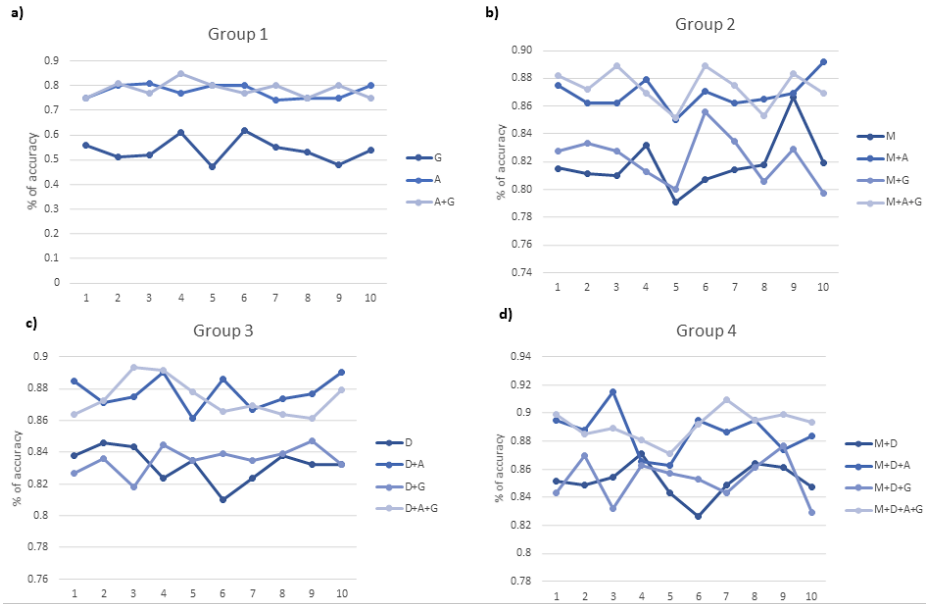

Fig. 3. Comparison of 10-fold cross-validation results: (a) Group 1, (b) Group 2, (c) Group 3, (d) Group 4.

Table 1. Evaluation Results For CVD AF Binary Classification.

HVFR Combination	Accuracy %	Precision %	Recall %	F1 %
Group 1 - Gender and Age				
Age (A)	77.60	79.60	77.60	77.70
Gender (G)	53.90	54.90	53.90	54.00
Age + Gender (A+G)	**78.30**	**80.50**	**78.30**	**78.50**
Group 2 - Motifs				
Motifs (M)	79.57	78.51	79.45	78.01
Motifs + Age (M+A)	**86.41**	**87.27**	**86.18**	**86.37**
Motifs + Gender (M+G)	80.17	77.52	82.0	79.22
Motifs + Age + Gender (M+A+G)	85.49	86.22	85.26	85.47
Group 3 - Discords				
Discords (S)	82.63	83.73	82.63	82.73
Discords + Age (S+A)	**87.78**	**88.23**	**87.78**	**87.78**
Discords + Gender (S+G)	82.42	83.80	82.42	82.54
Discords + Age + Gender (S+A+G)	87.50	87.85	87.50	87.50
Group 4 - Motifs and Discords				
Motifs + Discords (M+S)	85.56	85.99	85.56	85.58
Motifs + Discords + Age (M+S+A)	88.51	88.95	88.51	88.52
Motifs + Discords + Gender (M+S+G)	85.62	86.20	85.62	85.64
Motifs + Discords + Age + Gender (M+S+A+G)	**89.16**	**89.42**	**89.16**	**89.16**

6.2 Results

The average accuracy results of each TCV are presented in Figure 3 for each group. The average accuracy, precision, recall and F1 values are presented in Table 1. Figure 4 presents a graphical representation of the recorded F1 values from Table 1.

From Figure 4 it can firstly be observed that motifs, discords, age and gender when used on their own do not perform as well as when they are combined; thus supporting the motivation for a unifying representation. From the figure it can also been seen that using discords, without motifs, produces a better result than when using motifs, without discords, illustrating the advantage of using discords (not considered in, for example, [18] or [1]. It is interesting to note that including age does not make a significant difference despite the work presented in [20] that suggests it should do. Overall, the best result was obtained when a unifying HFVR was used that included motifs, discords, age and gender.

Thus, from the foregoing, it can be concluded that the idea of HFVR does indeed provide real benefits in the context of the CVD AF classification application; and, it is argued here, is likely to provide benefits with respect to other classification applications that feature heterogeneous data input. The reported evaluation has also provided empirical evidence that the proposed approach to HFVR generation, with respect to time series, operates well (produces good results).

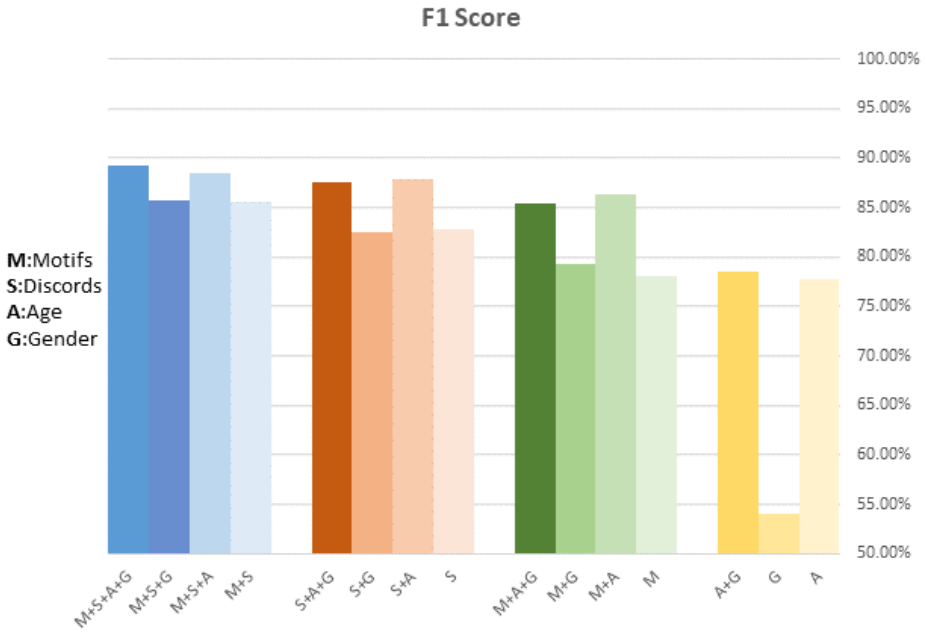

Fig. 4. Bar chart showing comparison of F1 values from Table 1.

7 Conclusion

This paper has reported on an investigation on the use of Homogeneous Feature Vector Representations (HFVRs) as a means allowing ML/DL over input data sets in different formats (heterogeneous data). The fundamental idea was, either directly or indirectly, to extract features from these data sets and incorporate them in to a unifying HFVR. In the case of tabular data, either numeric or categorical, this is a straight forwards process and can be achieved directly. In the case of other forms of data, such as time series data, this first requires the application of a feature extraction and selection process. The investigation was conducted using a CVD classification scenario; more specifically AF binary classification. The input here comprises time series ECG data and tabular data. For the time series data it was suggested that features to be extracted should be motifs and discords. It was further suggested that this be done using the CMP technique coupled with GMS. To support this idea experiments were conducted using four types of feature, motifs, discords, categorical data and numerical data. For the evaluation these four feature types were combined in different groupings. The results indicated that an HFVR made up of all four feature types produced the best results. Thus it was concluded that the idea of HFVR did indeed provide real classification benefits (at least in the context AF classification) and that the proposed approach to HFVR generation, with respect to time series, was a good one. For future work the authors intend to consider the most appropriate way of extracting features from images for inclusion in a unifying HVFR, and to apply the idea to alternative application domains.

References

1. Aldosari, H., Coenen, F., Lip, G., Zheng, Y.: Motif based feature vectors: Towards a homogeneous data representation for cardiovascular diseases classification. In: Proceedings of the The 23rd International Conference on Big Data Analytics and Knowledge Discovery (DaWaK'21) (2021)
2. Cabrera, D., Sancho, F., Li, C., Cerrada, M., Sánchez, R.V., Pacheco, F., Oliveira, J.V.: Automatic feature extraction of time-series applied to fault severity assessment of helical gearbox in stationary and non-stationary speed operation. Applied Soft Computing **58**, 53–64 (2017)
3. Christov, I., Krasteva, V., I. Simova, T.N., Schmid, R.: Multi-parametric analysis for atrial fibrillation classification in ecg. In: IEEE Computing in Cardiology (CinC'17). pp. 1–4 (2017)
4. Coyle, D., Prasad, G., McGinnity, T.M.: A time-series prediction approach for feature extraction in a brain-computer interface. IEEE Transactions on Neural Systems and Rehabilitation Engineering **13**(4), 461–467 (2005)
5. Das, M.K., Ari, S.: ECG beats classification using mixture of features. International Scholarly Research Notices **2014** (2014)
6. Dau, H.A., Keogh, E.: Matrix profile V: A generic technique to incorporate domain knowledge into motif discovery. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 125–134 (2017)
7. Ding, S., Du, M., Sun, T., Xu, X., Xue, Y.: An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood. Knowledge-Based Systems **133**, 294–313 (2017)
8. Golinko, E., Sonderman, T., Zhu, X.: Cnfl: categorical to numerical feature learning for clustering and classification. In: 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC). pp. 585–594. IEEE (2017)
9. Inoue, H., Atarashi, H., KenOkumura, . . ., Chishaki, A.: Impact of gender on the prognosis of patients with nonvalvular atrial fibrillation. American Journal of Cardiology **113**(6), 957–962 (2014)
10. Jain, A., Jain, V.: Voting ensemble classifier for sentiment analysis. In: Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020. pp. 255–261. Springer Singapore (2021)
11. Jovic, A., Bogunovic, N.: Feature extraction for ecg time-series mining based on chaos theory. In: Proceeedings 29th International Conference on Information Technology Interfaces. pp. 63–68 (2007)
12. Keogh, E.J., Lin, J., Fu, A.: Hot sax: Efficiently finding the most unusual time series subsequence. In: proceedings of the 5th IEEE International Conference on Data Minini (ICDM 2005). pp. 226–233 (2005)
13. Khalid, S., Khalil, T., Nasreen, S.: A survey of feature selection and feature extraction techniques in machine learning. In: Proceedings of the Science and Information Conference (SAI'14). pp. 372–378 (2014)
14. Kumar, D., Batra, U.: Breast cancer histopathology image classification using soft voting classifier. In: Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020. pp. 619–631. Springer Singapore (2021)
15. Li, P., Wang, Y., He, J., Wang, L., Tian, Y., Zhou, T., Li, T., Li, J.: High-performance personalized heartbeat classification model for long-term ECG signal. IEEE Transactions on Biomedical Engineering **64**(1), 78–86 (2016)

16. Lip, G., Fauchier, L., Freedman, S., Van Gelder, I., Natale, A., Gianni, C., Nattel, S., Potpara, T., Rienstra, M., Tse, H., Lane, D.: Atrial fibrillation. Nat Rev Dis Primers **31**, 16016 (2016)
17. Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., et al.: An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. Journal of Medical Imaging and Health Informatics **8**(7), 1368–1373 (2018)
18. 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: Proceedings of the Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support (2013)
19. Mueen, A., Keogh, E.J., at Chinese Academy of Sciences Qiang Zhu, Q.Z., Cash, S.: Exact discovery of time series motifs. In: Proceedings of the SIAM International Conference on Data Mining (SDM'09). pp. 473–484 (2009)
20. Naderi, S., Wang, Y., Miller, A.L., Rodriguez, F., K.Chung, M., Radford, M.J., M.Foody, J.: The impact of age on the epidemiology of atrial fibrillation hospitalizations. American Journal of Medicine **127**(2), 158.e1–158.e7 (2014)
21. Nady, S., Moness, M., Massoud, M., Gharieb, R.: Combining continuous wavelet transform and teager-kaiser energy operator for ECG arrhythmia detection. In: 8th Cairo International Biomedical Engineering Conference (CIBEC). pp. 76–79. IEEE (2016)
22. Padmavathi, S., Ramanujam, E.: Naïve bayes classifier for ECG abnormalities using multivariate maximal time series motif. Procedia Computer Science **47**, 222–228 (2015)
23. Sánchez-Cauce, R., Pérez-Martín, J., Luque, M.: Multi-input convolutional neural network for breast cancer detection using thermal images and clinical data. Computer Methods and Programs in Biomedicine **204**, 106045 (2021)
24. Sun, Y., Zhu, L., Wang, G., Zhao, F.: Multi-input convolutional neural network for flower grading. Journal of Electrical and Computer Engineering **2017** (2017)
25. Ventura, G., Benvenuti, E.: Advances in Discretization Methods: Discontinuities, Virtual Elements, Fictitious Domain Methods, vol. 12. Springer (2016)
26. Wang, X., Smith, K., Hyndman, R.: Characteristic-based clustering for time series data **13**, 35–364 (2006)
27. Wankhedkar, R., Jain, S.K.: Motif discovery and anomaly detection in an ECG using matrix profile. In: Progress in Advanced Computing and Intelligent Engineering, pp. 88–95. Springer (2021)
28. Yeh, C.C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Silva, D.F., Mueen, A., Keogh, E.: Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In: IEEE 16th International Conference on Data Mining (ICDM). pp. 1317–1322. IEEE (2016)
29. Zhao, Z., Särkkä, S., Rad, A.B.: Spectro-temporal ecg analysis for atrial fibrillation. In: proceedings of the 28th International Workshop on Machine Learning for Signal Processing (MLSP'18) (2018)
30. Zhu, Y., Yeh, C.C.M., Zimmerman, Z., Kamgar, K., Keogh, E.: Matrix profile XI: SCRIMP++: time series motif discovery at interactive speeds. In: IEEE International Conference on Data Mining (ICDM). pp. 837–846. IEEE (2018)