King's Research Portal

DOI:

[10.1109/TBME.2017.2785442](https://doi.org/10.1109/TBME.2017.2785442)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Alwan, Y., Cvetkovic, Z., & Curtis, M. J. (2018). Methods for Improved Discrimination between Ventricular Fibrillation and Tachycardia. *IEEE Transactions on Biomedical Engineering*, 65(10), 2143-2151. <https://doi.org/10.1109/TBME.2017.2785442>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Methods for Improved Discrimination between Ventricular Fibrillation and Tachycardia

Yaqub Alwan, Zoran Cvetković, *Senior Member, IEEE*, Michael J. Curtis

Abstract—Differentiating between ventricular tachycardia and ventricular fibrillation in clinical and preclinical research is based on subjective definitions that have yet to be validated using objective criteria. This is partly due to shortcomings in the discrimination ability of current objective approaches, typified by the algorithms that perform cardiac rhythm classification using low-dimensional feature representations of electrocardiogram (ECG) signals. These identify ventricular tachyarrhythmias, but do not discriminate between ventricular tachycardia and ventricular fibrillation. In order to address this limitation, we have tested the utility of high dimensional feature vectors, in particular, magnitude spectra and classifier ensembles that take into account local context information from ECG signals. Using these approaches we categorised rhythms into three classes: ventricular tachycardia, ventricular fibrillation and any other possible rhythm, defined here as ‘non-ventricular rhythms.’ The high-dimensional spectral features achieved a substantial improvement in the discrimination between ventricular tachycardia and ventricular fibrillation, but exhibited a decreased sensitivity to non-ventricular rhythms. In order to deal with the reduced sensitivity for detection of non-ventricular rhythms, methods were elaborated for combining the strengths of different feature spaces, and this substantially improved the identification sensitivities of all three classes.

Index Terms—Cardiac arrhythmias, ventricular tachycardia, ventricular fibrillation, classification, ensemble methods.

I. INTRODUCTION

CARDIOVASCULAR disease is among the leading causes of death worldwide [1]. The prevalence of ventricular fibrillation (VF), which is unpredictable and lethal, renders it a major unmet therapeutic need [2]. Clinical and preclinical research into VF has been disadvantaged by the subjective and inconsistent diagnostic criteria for discriminating between VF and ventricular tachycardia (VT) [3]–[5]. It is important to note that VT and VF are not always unequivocal, since VT is not always monomorphic, and a range of transient and persistent forms of polymorphic VT can occur. Discriminating between them and VF is not only challenging, it is important since although DC shock may be applied to all, in the longer term, including in patients in whom a defibrillator is not an option, appropriate intervention for some forms of polymorphic VT is different from that appropriate for VF [3]. When ventricular tachyarrhythmias are studied, in clinical cardiology, and in basic and clinical research, the basis for

labelling an arrhythmia as VT versus VF versus other rhythms is not standardized and can vary highly from study to study, with the most comprehensive system of classification based entirely on linguistic formalization of subjective criteria [3], rather than in terms of objective numerical criteria that can be used for automated detection and discrimination between VT and VF. Having such computational methods for arrhythmia discrimination is important because it is essential that a patient, a clinical event, and experimental data can be classified accurately, based on objective criteria.

The issue has spurred a large volume of research aiming to find features that can be derived from segments of ECG and used for reliable detection and classification of episodes of ventricular arrhythmias. However, none of these detection methods has become established in medical arrhythmia research. Many of these approaches report very high accuracy, however a more detailed reading often reveals methodological problems in the manner in which the accuracy of the proposed algorithms is evaluated. Rhythm examples which are difficult to classify are often excluded from consideration, or grouped into a separate category that has no medical or physiological basis. Another common methodological problem is the lack of proper separation between training and test data sets. Furthermore, frequently private databases are used, which makes it impossible to reproduce reported results.

Recently, two extensive studies were conducted to assess the ability of previously proposed features to detect and classify ventricular arrhythmias [6], [7]. The particular tasks that were investigated were the classification of non-ventricular rhythms (NVR) (including all rhythms that are not VT or VF) versus VT and VF combined [6], [7], and the classification of non-VF (NVR and VT) versus VF [7]. The effectiveness of the considered features and their combinations were ranked according to their performance on these tasks. Whereas the discrimination between VT and VF was not considered explicitly, in [7] it was observed that many confusions between VT and VF occur within the investigated tasks and it was concluded that discrimination between VT and VF is still very challenging.

Building on the work in [6], [7] the present study aims to: 1) evaluate previously proposed features in the context of multiclass classification between NVR, VT and VF, since this is more relevant than the two-class scenarios considered hitherto, and 2) to develop a methodology for improved discrimination between VT and VF in the context of three-class classification.

Support vector machines (SVMs) were used as baseline classifiers as they aim to establish maximum margin boundaries between classes and are particularly suitable for problems

This work was done while Yaqub Alwan was with the Department of Informatics, King’s College London. He is now with Sony Computer Entertainment Europe (e-mail: yaqub_alwan@scee.net). Zoran Cvetković is with the Department of Informatics, King’s College London, Strand, London, WC2R 2LS, United Kingdom (e-mail: zoran.cvetkovic@kcl.ac.uk). Michael Curtis is with the Cardiovascular Division, King’s College London, St Thomas’ Hospital, London, SE1 7EH, United Kingdom (e-mail: michael.curtis@kcl.ac.uk).

with relatively sparse training data. Since it was established in [7] that previously proposed low-dimensional features and their combinations do not provide for accurate discrimination between VT and VF, we undertook to use high-dimensional magnitude spectra features that retain more information about underlying ECG signals and demonstrate that they offer a considerably improved VT vs VF discrimination ability, even over ECG segments as short as 1 s. However, these features exhibit a lower sensitivity to NVR rhythms, therefore we explored methods for combining them with state-of-the-art features to improve sensitivity to all three classes. Finally, we explored ensemble methods, including stacked generalisation, to further improve diagnostic accuracy by exploiting context information via combining decisions made on consecutive ECG segments. As a result, improved sensitivity for all classes was achieved, on average close to 20%, with the most substantial improvement achieved with VT, where the sensitivity was increased by approximately 140%.

The contributions of this study are thus: 1) rigorous and extensive evaluation of state-of-the-art features in multiclass settings, 2) higher-dimensional feature spaces for improved discrimination between VT and VF, 3) ensemble methods for exploiting local context information in classification of cardiac arrhythmias, and 4) improved accuracy of cardiac arrhythmia detection and classification. This work was presented in part at the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [8].

II. METHODS FOR CLASSIFICATION

A. Representation Spaces

The relative ability of state-of-the-art features to discriminate between different arrhythmias has been recently assessed in two extensive studies using a genetic algorithm [6] and combined feature selection weighting [7]. For NVR vs arrhythmia classification two features were recommended in [6], whilst the assessment for non-VF vs VF classification revealed that the features become less relevant after the top six [7]. Therefore, for comparison here eight previously proposed features are used: the two features recommended in [6] and the top six from [7]. Given an ECG segment $\underline{x} = (x[0], x[1], \dots, x[N-1])$ consisting of N time samples, these features are:

- 1) **VF filter leakage** [9] – the residual energy of $\underline{x}$ after an adaptive bandstop filter, centred at the mean frequency of $\underline{x}$, is applied.
- 2) **Count 2** [10] – the number of samples of $\underline{x}$ between its mean and maximal values after a bandpass filter is applied.
- 3) **Threshold crossing sample count** [11] – the number of samples of $\underline{x}$ above an adaptive threshold.
- 4) **The sample entropy** [12] – a measure of self similarity in $\underline{x}$.
- 5) **Spectral parameter M** [13] – a normalised first moment of the spectrum of $\underline{x}$.
- 6) **Spectral parameter $A2$** [13] – a measure of the concentration of the spectrum of $\underline{x}$ around its main peak.
- 7) **PST** [14] – the number of unique value-pairs in a quantised phase space formed by $x[n]$ and $x[n+T]$, where T corresponds to 0.5 s.

- 8) **PSH** [15] – the number of unique value-pairs in a quantised phase space formed by $\underline{x}$ and its Hilbert transform.

In the present study, these features were used to form an 8-dimensional reference feature vector, referred to as *Heur8*.

These eight features, and other features proposed for cardiac rhythm classification so far, reflect heuristic observations about what could be properties of ECG waveforms that distinguish one rhythm from another. These features summarise ECG waveforms using a small set of numbers, most often just one or two. There is, however, no guarantee that this waveform summarisation does not discard information that is relevant for the discrimination between VT and VF. In fact, as observed in [7] and confirmed by our experiments reported in Section IV, it is extremely difficult to distinguish VT from VF in these low-dimensional spaces. In an attempt to improve the ability to discriminate between these two types of ventricular arrhythmias, we explored classification using raw ECG segments $\underline{x}$ that avoids any information loss. A problem associated with classification using raw waveforms is that a given ECG signal and its version shifted in time may exhibit a large Euclidean distance which makes it very challenging for automatic algorithms to recognise that such two waveforms belong to the same class. We therefore sought to achieve invariance to time shifts by considering magnitude spectra of waveform segments $\underline{x}$, since a waveform and its delayed version have the same magnitude spectrum. In particular, for an ECG segment $\underline{x}$ of length N we considered its magnitude spectrum vector $\underline{X} = (|X[1]|, |X[2]|, \dots, |X[N/2]|)$, which is the vector of the magnitudes of the first half of the discrete Fourier transform coefficients

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi}{N} kn}, k = 0, 1, \dots, N-1, \quad (1)$$

of $\underline{x}$. Note that since ECG signals are real, the Fourier transform satisfies the conjugate symmetry, $X[N-k] = X^*[k]$, so discarding coefficients $X[k]$, $k = N/2 + 1, \dots, N-1$ does not incur any information loss, and $X[0]$ is discarded since it is the mean of $\underline{x}$ and all records are normalised to zero mean. Such feature vectors $\underline{X}$ are here referred to as *Spectra*. We systematically compared the discrimination ability of Spectra features, $\underline{X}$, with complete underlying ECG waveforms, $\underline{x}$, and achieved better results using Spectra, which is probably because their shift-invariance facilitates learning of boundaries between different classes [16]. Therefore, in this study we undertook to use Spectra features and report results on ventricular rhythm classification using these features in combination with Heur8 features. It is important to note that the motivation behind introducing Spectra features was not to identify frequencies which are most informative for rhythm classification, but to obtain time-invariant feature vectors which preserve the most information about the underlying ECG waveforms.

Magnitude spectra have been previously considered in the literature in the context of cardiac rhythm classification. In [17] the authors undertook to extract 250 ms of QRS complexes and use magnitudes of five low-frequency components of their spectra for rhythm classification. Since according to

the leading guidance the QRS complex is not identifiable in VF by definition [3], [18], we undertook not to extract QRS complexes but to consider magnitude spectra of longer sliding windows of ECG signals, typically 1 s long. More crucially, rather than relying on any heuristics on what would be the most appropriate frequency bands, we included all the information by considering complete magnitude spectra, which yields representations of much higher dimensions than considered previously. For example, for the sampling rate of 100 Hz, which is used in this study, 1 s segments $\underline{x}$ of ECG are vectors of dimension 100, hence the corresponding Spectra features $\underline{X}$ are 50-dimensional vectors; the dimension increases linearly with segment duration.

B. Multi-class classification with SVMs

For the standard two-class SVM, with output labels in $\{1, -1\}$, the distance of a data point, represented by its feature vector $\underline{X}$, from the decision boundary is given by

$$f(\underline{X}) = \sum_i \alpha_i l_i K(\underline{X}, \underline{X}_i) + b, \quad (2)$$

where α_i and b are optimised by the training algorithm, $\underline{X}_i$ and l_i are the i^{th} training data point and its label, respectively, and $K(\cdot, \cdot)$ is some positive definite kernel function. The classification decision is made based on the sign of the distance of $\underline{X}$ from the boundary, $C(\underline{X}) = \text{sgn}(f(\underline{X}))$, where sgn is the sign function. A commonly used kernel is the radial basis function (RBF) kernel, $K(\underline{X}, \underline{Y}) = e^{-\gamma \|\underline{X} - \underline{Y}\|^2}$, $\gamma \in \mathbb{R}^+$. We systematically compared the RBF kernel to polynomial kernels of different orders, and this achieved the highest accuracy. Hence this kernel was used in all experiments reported here.

Multiclass discrimination is achieved by combining binary SVMs via the error correcting output codes (ECOC) approach [19], [20]. To that end, L binary classifiers are trained, each distinguishing between two groups of M classes. The groups are specified by elements $\mathbf{W}_{mn} \in \{1, 0, -1\}$, of an $M \times L$ coding matrix $\mathbf{W}$, where $\mathbf{W}_{mn}$ is the label of the class m in the n -th binary classification task, with the 0 label assigned to classes which are not included for training of the particular binary SVM. In the three-class case, different groupings can be formed from the one-vs-one and one-vs-all tasks only. This gives $L = 6$ binary classifiers, which were all used in this study, with the corresponding matrix $\mathbf{W}$ being

$$\mathbf{W} = \begin{bmatrix} 1 & 1 & -1 & 1 & 1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 1 \\ -1 & 1 & 1 & -1 & 0 & -1 \end{bmatrix}. \quad (3)$$

Using this coding matrix, the class assignment rule is then of the form

$$C(\underline{X}) = \arg \min_m \sum_{n=1}^L \chi(\mathbf{W}_{mn} f^n(\underline{X})), \quad (4)$$

where f^n is the decision function of the n^{th} classifier, and $\chi(z)$ is a loss decoding function. Loss functions explored in this study were: $\max(1 - z, 0)$ – hinge loss, $0.5(1 - \text{sgn}(z))$ – Hamming loss, e^{-z} – exponential loss, $-z$ – linear loss.

III. ENSEMBLE METHODS AND FEATURE COMBINING

A. Temporal ensembles

Classification accuracy can often be improved by combining decisions of ensembles of classifiers [21]. Local context information is important in rhythm classification and we exploited it by forming ensembles of classifiers temporally. For that purpose, consider vectors of decision values produced over T consecutive, possibly overlapping, ECG segments $\underline{x}_t$:

$$\underline{f}^n = (f_1^n, f_2^n, \dots, f_T^n). \quad (5)$$

where $f_t^n = f^n(\underline{X}_t)$ is the decision value of the n -th binary SVM corresponding to the feature vector $\underline{X}_t$ derived from the ECG segment $\underline{x}_t$. Then, using an aggregation function $A : \mathbb{R}^T \rightarrow \mathbb{R}$ and combining it with (4) the decision rule becomes

$$C(\underline{X}_1, \dots, \underline{X}_T) = \arg \min_m \sum_{n=1}^L \chi(\mathbf{W}_{mn} A(\underline{f}^n)). \quad (6)$$

Aggregation functions $A(\underline{f})$ which were explored in this study were: $\text{mean}(\underline{f})$ – mean, $\text{median}(\underline{f})$ – median, $\text{mode}(\text{sgn}(\underline{f}))$ – majority vote, $\underline{f}_{\arg \max_t |f_t|}$ – maximum absolute value.

Alternatively, temporal ensembles may be formed from the loss values after decoding rather than from the SVM outputs. For that purpose a matrix of loss values is formed,

$$\mathbf{L}_{mt} = \sum_{n=1}^L \chi(\mathbf{W}_{mn} f^n(\underline{X}_t)), \quad (7)$$

$m = 1, 2, \dots, M, \quad t = 0, 1, \dots, T.$

Then, the decision can be made according to

$$C(\underline{X}_1, \underline{X}_2, \dots, \underline{X}_T) = \arg \min_m B((\mathbf{L}_{m1}, \mathbf{L}_{m2}, \dots, \mathbf{L}_{mT})), \quad (8)$$

where B is an aggregation function. Specific functions $B(\underline{f})$ which were explored in this study were: $\text{mean}(\underline{f})$, $\text{median}(\underline{f})$, $\text{min}(\underline{f})$, whilst in the case of majority voting the decision was made according to $C(\underline{X}_1, \underline{X}_2, \dots, \underline{X}_T) = \text{mode}((c_1, c_2, \dots, c_T))$, where $c_t = \arg \min_m \mathbf{L}_{mt}$.

This family of schemes for making decisions incorporating temporal information from SVMs with ECOCs is referred here to as local context ensembles (LCEs). Note that LCE decoding can be performed either forwards or backwards temporally. When it is performed forwards, a latency is introduced. In order to reduce decision latency, only past observations were considered.

B. Stacking with SVM

The LCE approach described in III-A is conceptually and computationally simple, and it is often very effective. There are however many other well principled ways to combine the decisions of the ensembles of the binary classifiers. Here we additionally used stacked generalisation, which is described next.

Consider a training set $\mathcal{X}$ of feature vectors $\underline{X}$ of dimension P . Then, for a K -fold cross validation procedure where $\mathcal{X}$ is

split into K non-overlapping sets $\mathcal{X}^k$, let the training data set for the k^{th} -fold be $\hat{\mathcal{X}}^k = \mathcal{X} \setminus \mathcal{X}^k$. With a learning function $\mathcal{L}$, a model is learned with each training data subset,

$$\underline{F}_k = \mathcal{L}(\hat{\mathcal{X}}^k), \quad k = 1, \dots, K, \quad (9)$$

where $\underline{F}_k : \mathbb{R}^P \rightarrow \mathbb{R}^L$, i.e. $\underline{F}_k$ outputs a row vector of dimension L given an input feature vector of dimension P . In the case of the binary SVM classifiers considered in the previous subsection, $\underline{F}_k(\underline{X}) = (f_k^1(\underline{X}), f_k^2(\underline{X}), \dots, f_k^L(\underline{X}))$, where f_k^n is the decision function of the n^{th} binary classifier, learned in the k^{th} fold.

Then a new training data set $\mathcal{X}'$ is formed from concatenated outputs of models learned on individual folds, and evaluated at each training data point $\underline{X}_i$, i.e.

$$\underline{X}'_i = [\underline{F}_1(\underline{X}_i), \dots, \underline{F}_K(\underline{X}_i)], \quad \forall i. \quad (10)$$

Different parts of the data set are unseen by different models, and it is this which allows the final classifier to learn about the bias of the individual classifiers. Then, in order to capture local context information too, each data point is further augmented by including the same representation corresponding to T consecutive ECG segments:

$$\underline{X}'_i = [\underline{X}'_{i1}, \underline{X}'_{i2}, \dots, \underline{X}'_{iT}], \quad \forall i. \quad (11)$$

A meta-level classifier is then trained on this set, and classification is performed using feature vectors which are created in the same manner.

The resulting dimension of such a representation can be excessively high; e.g. in the case of an error correcting code for 3 class SVMs with $L = 6$ binary classifiers, 5-fold cross validation, and 15 temporal outputs, the resulting feature dimension is $6 \cdot 5 \cdot 15 = 450$. This is problematic as there may not be enough data to learn from in such high dimensions. Two strategies were used in this study for alleviating this dimension explosion.

- A) Use of fewer folds for cross-validation. The minimum reasonable number is 3 folds, since the amount of training data available to the lower level SVM classifiers is reduced (5 fold uses 80% of the data, while 3 fold uses 66% of the data). The reduction in dimension is 40% compared to 5-fold cross validation.
- B) Not using cross-validation at all. Instead, (ab)using the fact that categories are subsampled to obtain balanced SVM training data across categories, the remainder of data can be used as ‘‘held out’’ data for the stacking process.

Due to the expected high dimension of the final representation $\underline{X}'$, only the simple linear discriminant analysis (LDA) classifier was considered for the meta-level classification. Preliminary experiments testing a variety of linear and non-linear classifiers found no significant benefits of using more sophisticated classifiers at this level, which could be due to the fact that more sophisticated classifiers might need much more training data to be fully effective. Additionally, LDA is a computationally efficient option, allowing development of many models with different amount of temporal context in order to assess the impact of its inclusion.

C. Combining Strengths of Feature Sets

Results reported in Section IV show that whilst Spectra features improve the discrimination between VT and VF, Heur8 features exhibit higher sensitivity to NVR rhythms. In this section we describe three strategies that were used to leverage strengths of both representations.

1) *Concatenated features representation*: A common approach to combining strengths of different feature sets is their concatenation. To that end, given an ECG segment, $\underline{x}$, we considered feature vectors

$$\underline{X} = [\underline{X}_H \quad \underline{X}_S] \quad (12)$$

where $\underline{X}_H$ and $\underline{X}_S$ are Heur8 and Spectra features, respectively. Then, $\underline{X}$ can be substituted directly into (9), (10) and (11) in order to perform temporal stacking.

2) *Stacking over multiple feature spaces*: Different feature sets can be combined at the level of stacked generalisation. For that purpose, in the process of stacked generalisation, for each fold, separate classifiers were trained for each feature set and then a new training set was created as

$$\underline{X}'_i = [\underline{F}_{H1}(\underline{X}_{Hi}), \dots, \underline{F}_{HK}(\underline{X}_{Hi}), \underline{F}_{S1}(\underline{X}_{Si}), \dots, \underline{F}_{SK}(\underline{X}_{Si})] \quad (13)$$

where $\underline{F}_{Hk}$ and $\underline{F}_{Sk}$ are the k^{th} fold classifiers, trained on Heur8 features and Spectra features, respectively, and $\underline{X}_{Hi}$ and $\underline{X}_{Si}$ are Heur8 and Spectra features, respectively, derived from the same ECG segment $\underline{x}_i$. Then, $\underline{X}'_i$ was substituted into (11) in order to obtain temporally stacked data for training, and stacked generalisation was performed using these features as described in Section III-B

3) *A hierarchical approach*: Strengths of different features can be combined using a hierarchical approach to decision making too [22], [23]. To that end, consider making decisions on a given ECG segment simultaneously by classifiers C_1 and C_2 operating on different sets of features. Let C_1 be the primary classifier. If a NVR decision is taken by the primary classifier, then take that decision, otherwise take the decision of the secondary classifier C_2 :

$$C(\underline{X}_p, \underline{X}_s) = \begin{cases} C_1(\underline{X}_p), & C_1(\underline{X}_p) = \text{NVR} \\ C_2(\underline{X}_s), & C_1(\underline{X}_p) \neq \text{NVR} \end{cases} \quad (14)$$

where $\underline{X}_p$ and $\underline{X}_s$ are the primary and the secondary feature vectors, respectively, derived from the same ECG segment $\underline{x}$. C_1 and C_2 can be composed in many ways, by varying both the classification algorithms and feature spaces. In Section IV we report results where:

- $\underline{X}_p$ are Heur8 features, and $\underline{X}_s$ are concatenated Spectra and Heur8 features,
- classifiers C_1 and C_2 are local context ensemble classifiers, as described in Section III-A, or stacked classifiers described in Section III-B,
- both C_1 and C_2 make three-class, i.e. NVR, VT, VF, decisions.

TABLE I
DATABASES USED IN THIS STUDY.

Database	Number of records	Record length
AHADB	154	35 min
VFDB	22	35 min
EDB	90	120 min
CUDB	35	8.5 min
MITDB	48	30 min

TABLE II
THE TOTAL DURATION OF THE THREE RHYTHM EXAMPLES USED IN THE STUDY.

Rhythm type	Duration
NVR	144,502 s
VT	6,477 s
VF	16,727 s

IV. EXPERIMENTS AND RESULTS

A. Data and preprocessing

Data were taken from Physiobank [24] and the extended American Heart Association Database (AHADB). The databases from Physiobank include the European ST-T Database (EDB) [25], the Creighton University Ventricular Tachyarrhythmia Database (CUDB) [26], the MIT-BIH Arrhythmia Database (MITDB) [27], and the MIT-BIH Malignant Ventricular Arrhythmia Database (VFDB) [28]. The number of records in each database and their duration is shown in Table I. From these databases only records containing examples of VT or VF were used, which is 98 records in total. Any records containing annotations for so-called ventricular flutter were excluded, due to ambiguity about which category these rhythms belong to. Ventricular flutter does not have a commonly agreed definition and moreover it is not recognised by the leading guidance for the study of ventricular arrhythmias [3]. This resulted in 91 out of 98 possible records being used. The total duration of the three investigated rhythm types, NVR, VT and VF, in these 91 records is shown in Table II.

All records were resampled to 100 Hz sampling rate and processed by a 0.5 Hz high pass FIR filter for removal of baseline wander. The ECG records were further normalised to unit sample variance.

B. Nonparametric error estimation

Classification accuracy estimates were obtained using bootstrap resampling (sampling without replacement) with 80% of ECG records held for training and the remaining 20% used for accuracy assessment, ensuring that records used for training and records used for testing were completely separate. This was repeated 200 times with different randomised sampling of the test records. We were unable to ascertain whether each of the 91 used records came from a different patient, so it is possible that in some of the resamplings some records used for training and some records used for testing came from same patients. Although there is no evidence that the morphology

TABLE III
THE EXPERIMENTAL PARAMETERS AND VALUES FOR EXPERIMENTS CONDUCTED IN THIS STUDY, WITH ACRONYMS FOR REFERENCING METHODS IN TABLES AND FIGURES.

Parameter	Values
Observation length	1 second with overlap, each segment shifted by 0.25 seconds
Classifiers	RBF SVM
Ensembles	LCE, temporal stacked generalisation without cross validation (SG), temporal stacked generalisation by 3-fold cross validation (SG3CV), hierarchical (H- prefix, followed by primary then secondary classifiers)
Representations	Spectra, Heur8, Spectra and Heur8 concatenated (S+H8), Spectra and Heur8 trained separately and stacked together (S/H8) (LCE is not applicable in this case)

of arrhythmias are unique to a patient, that could potentially make the accuracy higher than if each record comes from a different patient. Nevertheless, all methods and feature sets were compared on the same data, ensuring that the conclusions of the study about their relative merits are valid. As the base ECG segment length is set to 1 s (see Table III), in each run the accuracy was assessed using approximately 28,900, 1,290, and 3,340 examples of NVR, VT, and VF, respectively, which is 20% of the total duration of the three rhythms as shown in Table II. Where hyper-parameter estimation needed to be performed, such as with SVM cost parameter C and RBF kernel parameter γ , five fold cross-validation over the training records was used. Cross validation folds were formed by records, rather than by ECG segments. The parameters were selected according to the best performing on average across all five folds. The selected parameters were used for a final training pass with the 80% training data records, leaving the remaining records entirely unseen for testing.

For two-class classification, typically sensitivity and specificity metrics are reported. However, since sensitivity is accuracy of the positive class, and specificity is accuracy of the negative class, we have determined the sensitivity of each given category. An additional metric that is reported here is the average of individual sensitivities, sn_s , across the set of classes S ,

$$Acc_{bal} = \frac{1}{|S|} \sum_{s \in S} sn_s. \quad (15)$$

C. Experimental results

Experiments were conducted using the same 200 bootstrap resamples for all methods tested. All experimental parameters that were tested are listed in TABLE III. For LCEs, the possible parameters are listed in TABLE IV. Finally, for the LCE and stacking ensembles and representation combinations, the parameters which gave best classification results, according to median Acc_{bal} values, are reported in TABLE V. Fig. 1 shows Acc_{bal} and per category sensitivity distributions for each of Heur8, Spectra and S+H8 representations with their respective best parameters (as shown in TABLE V) for LCE, SG and SG3CV ensembles. Fig. 2 shows Acc_{bal} and per category sensitivity distributions for a selection of benchmark methods, the best temporal ensembles and the best

TABLE IV

THE PARAMETERS FOR LCE DECODING ARE LISTED. THE TOTAL NUMBER OF DIFFERENT COMBINATIONS OF THESE PARAMETERS RESULTS IN 224 DIFFERENT TYPES OF LCE

Parameter	Values and ordering
Temporal context	2 seconds, up to 8 seconds, i.e. 5 segments, up to 29 segments
Loss function	hinge, linear, exponential, Hamming
Aggregation function	mean, median, majority vote, maximum absolute value/minimum value (depends on whether before or after loss decoding)
Aggregation level	Before loss decoding, after loss decoding

TABLE V

THE BEST PERFORMING PARAMETER COMBINATIONS ARE LISTED FOR EACH OF THE DIFFERENT ENSEMBLE TYPES BUILT WITH EACH REPRESENTATION SPACE

Representation	Ensemble	Best parameter combination
Spectra	LCE	5 s temporal aggregation, Hamming loss decoding, aggregation with minimum loss after loss decoding
Heur8	LCE	7 s temporal aggregation, exponential loss decoding, aggregation with mean loss after loss decoding
S+H8	LCE	8 s temporal aggregation, Hamming loss decoding, aggregation with median function before loss decoding
Spectra	SG	8 s temporal aggregation, stacking Hamming loss decoding values
Heur8	SG	8 s temporal aggregation, stacking without loss decoding
S+H8	SG	8 s temporal aggregation, stacking linear loss decoding values
Spectra	SG3CV	7 s temporal aggregation, stacking without loss decoding
Heur8	SG3CV	8 s temporal aggregation, stacking hinge loss decoding values
S+H8	SG3CV	7 s temporal aggregation, stacking without loss decoding

mixed representation/hierarchical ensembles; in most cases, variations in category sensitivities and Acc_{bal} scores due to various hierarchical constructions were small. For comparison, classification results obtained by using Heur8 features derived from 8 s and Spectra features derived from 1 s of non-overlapping ECG segments and simple multiclass RBF SVMs are also presented. Finally, average confusion matrices across all bootstrap resamples are shown in TABLE VI.

D. Discussion

An important observation to be made from the results presented in Table VI is that the average sensitivity of NVR vs VT vs VF classification using Heur8 features is relatively low, 62.3%, and that this is due to the lack of the ability of Heur8 features to discriminate accurately between VT and VF. The discrepancy between these results and some assessments previously published comes from differences in training and assessment methodologies, in that here records used for training are completely separate from those used in testing, and there is no pre-selection of data.

Fig. 1. Distributions of (a) Acc_{bal} and (b) per category sensitivities, across all bootstrap resamples of LCE, SG and SG3CV ensembles with Spectra, Heur8 and S+H8 representations, shown in each case for the best parameter combination by median Acc_{bal} as described in TABLE V

(a)

(b)

Fig. 2. Distributions of (a) Acc_{bal} and (b) per category sensitivities, across all bootstrap resamples of Heur8 and Spectra reference classifiers, the best performing Spectra temporal ensemble, the best performing concatenated features temporal ensemble, hierarchical classification via stacking Spectra and Heur8 separately, and hierarchical decision combining

TABLE VI
AVERAGE CONFUSION MATRICES OVER ALL BOOTSTRAP RESAMPLES FOR EACH REPRESENTATION SHOWN IN FIG. 2. ROWS ARE THE GROUND TRUTHS, AND COLUMNS ARE THE DIAGNOSES MADE.

Method	Acc_{bal} (%)	Truth	Diagnosed as		
			NVR%	VT%	VF%
Heur8 8s	62.3	NVR	92.4	4.8	2.9
		VT	14.1	21.0	64.9
		VF	8.3	18.1	73.6
Heur8 SG3CV	66.9	NVR	97.5	1.5	1.0
		VT	22.4	27.1	50.4
		VF	6.3	17.5	76.2
Spectra 1s	66.9	NVR	76.1	8.4	15.5
		VT	19.0	45.1	35.9
		VF	4.7	15.7	79.6
Spectra SG	70.7	NVR	83.9	5.3	10.8
		VT	22.4	49.1	28.5
		VF	3.1	17.6	79.2
S+H8 LCE	74.1	NVR	89.8	5.2	5.0
		VT	18.4	51.9	29.7
		VF	3.7	15.6	80.7
H-LCE Heur8 LCE S+H8	74.7	NVR	94.4	2.6	3.0
		VT	20.4	50.0	29.6
		VF	4.9	15.3	79.8
S/H8 SG3CV	71.1	NVR	97.7	1.1	1.2
		VT	23.4	35.2	41.5
		VF	7.9	11.6	80.5
H-SG Heur8 SG S+H8	73.5	NVR	96.2	1.6	2.2
		VT	22.3	48.9	28.8
		VF	4.8	19.6	75.5

Benefits of time ensembles can be observed by comparing the results obtained by using Heur8 features derived from segments of length 8 s and ensembles of Heur8 features derived from 1 s segments over 8 s intervals (the first two rows of Table VI): combining decisions made on shorter intervals rather than making a decision based on one feature vector derived on the equivalent time interval increased the average class sensitivity from 62.3% to 66.9%. Stacked generalisation over temporal feature ensembles improved also the accuracy of classification using Spectra features, from 66.9%, obtained with Spectra derived from 1 s long ECG segments, to 70.7%, achieved with ensembles of Spectra features derived from 1 s intervals and collected over 8 s windows (Table VI).

Further we found that using Spectra in place of Heur8 features increased the average sensitivity. In particular, Spectra features derived from only 1 s long ECG segments increased the average sensitivity from 62.3% to 66.9% compared to Heur8 features derived from 8 s long segments, and further improvement to 70.7% was achieved via temporal ensembles of Spectra features over the same 8 s intervals. Confusion matrices in Table VI reveal that this improvement was due to the higher ability of Spectra features to discriminate between VT and VF. However that came at the cost of a lower sensitivity to NVR. This sensitivity trade-off motivated considering the concatenation of the two feature sets.

Figure 1a illustrates improvements in average class sensitivity gained by Spectra and concatenated features compared to Heur8 features alone in the context of temporal ensembles. From Figure 1b it can be further observed that concatenated Spectra and Heur8 features with local context

ensembles achieved the highest sensitivities to VT and VF, even compared to more sophisticated stacking methods, with average class sensitivity of 74.1%, as detailed in row S+H8 LCE of Table VI. However, even for the concatenated features, the NVR sensitivity fell short compared with *Heur8* representation, with or without temporal ensembles. Hence, we investigated a hierarchical approach, as described in Section III-C3, where *Heur8* are used as primary features whilst the concatenated *Heur8* and *Spectra* features are used as the secondary features, and both classifiers use LCE decoding. This hierarchical classifier achieved the highest average sensitivity for the three classes, with the relative improvement of the average sensitivity compared to *Heur8* close to 20%, the most substantial improvement being for VT, which was close to 140%.

The effectiveness of different methods of aggregating decisions made on temporal feature ensembles is also illustrated in Figure 1a. The box plots shown in the figure suggest that in terms of average sensitivity to all three classes there is no benefit in using stacked generalisation rather than the simpler LCE method. However, Figure 1b reveals that in terms of the sensitivity to NVR, stacked generalisation gives better results than the LCE decoding for all feature sets. In fact the highest sensitivity to NVR, 97.7%, was achieved with stacking over the two feature spaces, as described in Section III-C2. The confusion matrix for this approach, shown in row S/H8 SG3CV of Table VI, reveals that this high sensitivity to NVR was obtained at the expense of a considerable drop in the sensitivity to VT. To achieve a better balance of sensitivities to different classes we introduced a hierarchical classifier with *Heur8* as the primary and concatenated *Heur8* and *Spectra* as the secondary feature set, with both classifiers having the stacked generalisation architecture as described in Section III-B. The confusion matrix for this approach is shown at the bottom of Table VI.

Average and per-class sensitivities of the discussed methods are summarised in Figure 2 using box plots to illustrate the distribution of results over 200 bootstrap resamples of the data sets. Figure 2a illustrates the improvements in average sensitivity made by replacing *Heur8* features by *Spectra*, and further through stacked generalisation, to local context ensembles of concatenated features. Figure 2b shows trade-offs between individual class sensitivities through which these improvements were achieved. Finally, sensitivities of methods S/H8 SG3CV, H-LCE *Heur8* LCE S+H8, and H-SG *Heur8* SG S+H8, are shown – these three methods achieved higher sensitivities for all three classes than plain *Heur8* features, and offered different trade-offs between individual class sensitivities.

The improvements in sensitivities to VT and VF achieved by using *Spectra* features suggests that the two arrhythmias are better separated in this high-dimensional representation space that avoids excessive heuristic information reduction. The accuracy of a classification algorithm in a certain feature space depends, however, not only on the separation of the classes in the space, but also on our ability to learn their distributions or boundaries. The amount of data points required to learn distributions generally increases exponentially in the

dimension of the feature space. However, even all publicly available curated databases of ventricular arrhythmias, as used in this study, provide very sparse training data for learning in dimensions close to the dimension of *Spectra* features. Nevertheless, with the SVM methodology, which aims to learn only class boundaries rather than complete distributions and is thus well suited for classification using sparse data, we achieved a substantial increase in the accuracy of the discrimination between VT and VF. This effect of the amount of training data is probably the reason why for the easier problem of the discrimination between NVR and arrhythmias, better results were consistently obtained using low-dimensional features. We anticipate that as more data become available it will be possible to achieve improvements in sensitivities for all classes using just high-dimensional features, possibly even raw ECG waveforms. The situation is analogous to what is presently happening with automatic speech recognition, where with the advent of deep learning and availability of large amounts of data it has become possible to achieve state-of-the-art results using raw acoustic waveforms of speech [29].

The availability of curated arrhythmia databases is a separate issue. The problem of ventricular arrhythmia classification was approached in the present study by the way of supervised learning. This means that rhythms were provided with ground truths, as a gold standard. The rhythm labels were used to estimate a mapping function from a feature space to a decision about which rhythm is being observed in the ECG. An aspect that requires consideration is the quality of ECG rhythm labelling. Given the difficulty of experts in reaching agreement on some examples of labelled ECGs [30], the existence of the Lambeth Conventions articles [3], [18] and a study where human experts had difficulty in classifying some of the evaluation traces [10], it is clear that one element of the problem which is critical to achieving success in building an objective rhythm detector is that we do not yet have reliable gold standards for either development or assessment of the algorithms. Of particular note is the fact that a more recent study [6] relabelled portions of NVRs as VT in the Creighton University Ventricular Tachyarrhythmia Database, acknowledging a critical problem with the gold standard that potentially limits the upper bound on achievable accuracy of automated diagnostics.

V. CONCLUSIONS

In past studies, many features have been proposed to provide discrimination in the ECG between some combinations of the categories considered in the present study, namely NVR, VT and VF. Two recent studies [6], [7] assessed combinations of developed features in their ability to discriminate between NVRs and arrhythmias, or between non-VF and VF. Whilst the discrimination between the three classes, or between VT and VF, were not directly considered, it was observed that many confusions between the arrhythmias occurred and it was concluded that the discrimination between them was a difficult problem [7].

In the present study, methods were developed to differentiate between NVR, VT and VF. In order to improve

the discrimination between VT and VF, a high-dimensional feature set was introduced aiming to retain more diagnostically relevant information from ECG signals. This alone improved average diagnostic accuracy but resulted in a decreased NVR sensitivity. In order to improve upon this, ensembles of consecutive decision outputs from the ECG were combined into a higher level decision process that captures temporal context. This resulted in a substantial improvement, but with NVR sensitivity still lower than that achieved by means of existing features. Therefore, methods were developed to combine the predictive powers of high-dimensional features and the reference features. The cumulative impact of all the contributions was to substantially reduce confusions between VT and VF whilst improving the sensitivity for identifying NVR.

The results obtained in this study indicate that there is merit in the elaboration of sequential methods for discriminating between different ventricular tachyarrhythmias. Our results further suggest that high-dimensional representations that avoid information loss could be the key towards better automatic classification of cardiac rhythms, and that the full potential of this approach will be attained as additional databases become available.

ACKNOWLEDGMENT

We would like to thank anonymous reviewers for their comments which helped us to significantly improve the presentation of the paper.

REFERENCES

- [1] World Health Organization, "WHO — the top 10 causes of death," <http://www.who.int/mediacentre/factsheets/fs310/en/>, 2014, [Online; accessed 09-December-2014].
- [2] T. Smith and M. Cain, "Sudden cardiac death: Epidemiologic and financial worldwide perspective," *Journal of Interventional Cardiac Electrophysiology*, vol. 17, no. 3, pp. 199–203, 2006.
- [3] M. J. Curtis *et al.*, "The Lambeth conventions (II): guidelines for the study of animal and human ventricular and supraventricular arrhythmias," *Pharmacology & Therapeutics*, vol. 139, no. 2, pp. 213–248, 2013.
- [4] F. Kette *et al.*, "What is ventricular tachycardia for an automated external defibrillator?" *Journal of Clinical & Experimental Cardiology*, vol. 5:285, 2014.
- [5] M. Madhavan and P. Friedman, "Optimal programming of implantable cardiac-defibrillators," *Circulation*, vol. 128, no. 6, pp. 659–672, 2013.
- [6] Q. Li *et al.*, "Ventricular fibrillation and tachycardia classification using a machine learning approach," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1607–1613, June 2014.
- [7] F. Alonso-Atienza *et al.*, "Detection of life-threatening arrhythmias using feature selection and support vector machines," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 832–840, March 2014.
- [8] Y. Alwan *et al.*, "Structured prediction for differentiating between normal rhythms, ventricular tachycardia, and ventricular fibrillation in the ECG," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, September 2015, pp. 310–314.
- [9] S. Kuo and R. Dillman, "Computer detection of ventricular fibrillation," in *Computers in Cardiology*, 1978, pp. 347–349.
- [10] I. Jekova and V. Krasteva, "Real time detection of ventricular fibrillation and tachycardia," *Physiological Measurement*, vol. 25, no. 5, pp. 1167–1178, Oct 2004.
- [11] M. Arafat *et al.*, "A simple time domain algorithm for the detection of ventricular fibrillation in electrocardiogram," *Signal, Image and Video Processing*, vol. 5, no. 1, pp. 1–10, 2011.
- [12] H. Li *et al.*, "Detecting ventricular fibrillation by fast algorithm of dynamic sample entropy," in *IEEE International Conference on Robotics and Biomimetics*, Dec 2009, pp. 1105–1110.
- [13] S. Barro *et al.*, "Algorithmic sequential decision-making in the frequency domain for life threatening ventricular arrhythmias and imitative artefacts: a diagnostic system," *Journal of Biomedical Engineering*, vol. 11, no. 4, pp. 320–328, 1989.
- [14] A. Amann *et al.*, "Detecting ventricular fibrillation by time-delay methods," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 1, pp. 174–177, Jan 2007.
- [15] —, "A new ventricular fibrillation detection algorithm for automated external defibrillators," in *Computers in Cardiology*, Sept 2005, pp. 559–562.
- [16] Y. Alwan *et al.*, "High-dimensional discriminant analysis of human cardiac arrhythmias," in *European Signal Processing Conference*, Sept 2013, pp. 1–5.
- [17] K. Minami *et al.*, "Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network," *IEEE Transactions on Biomedical Engineering*, vol. 46, no. 2, pp. 179–185, Feb 1999.
- [18] M. J. A. Walker *et al.*, "The Lambeth conventions: guidelines for the study of arrhythmias in ischaemia, infarction, and reperfusion," *Cardiovascular Research*, vol. 22, no. 7, pp. 447–455, 1988.
- [19] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of Artificial Intelligence Research*, vol. 2, pp. 263–286, 1995.
- [20] E. L. Allwein *et al.*, "Reducing multiclass to binary: a unifying approach for margin classifiers," *Journal of Machine Learning Research*, vol. 1, pp. 113–141, September 2001.
- [21] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.
- [22] J. C. Platt *et al.*, "Large margin DAGs for multiclass classification," in *Advances in Neural Information Processing Systems*. MIT Press, 2000, pp. 547–553.
- [23] R. Tibshirani and T. Hastie, "Margin trees for high-dimensional classification," *Journal of Machine Learning Research*, vol. 8, pp. 637–652, May 2007.
- [24] A. L. Goldberger *et al.*, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [25] A. Taddei *et al.*, "The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography," *European Heart Journal*, vol. 13, no. 9, pp. 1164–1172, 1992.
- [26] F. Nolle *et al.*, "CREI-GARD, a new concept in computerized arrhythmia monitoring systems," in *Computers in Cardiology*, vol. 13, 1986, pp. 515–518.
- [27] G. Moody and R. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, June 2001.
- [28] S. Greenwald, "Development and analysis of a ventricular fibrillation detector," Master's thesis, Department of Electrical Engineering and Computer Science, MIT, 1986.
- [29] T. Sainath *et al.*, "Learning the speech front-end with raw waveform CLDNNs," in *Proc. Interspeech*, 2015.
- [30] R. Clayton *et al.*, "Self-terminating ventricular tachyarrhythmias – a diagnostic dilemma?" *The Lancet*, vol. 341, no. 8837, pp. 93–95, 1993.