Classifying Left Ventricular Hypertrophy from Extracted Morphological Electrocardiogram Biomarkers using Random Forest

Hafiz Naderi[1], Julia Ramírez[1,2], Stefan van Duijvenboden[1], Esmeralda Ruiz Pujadas[3], Lin Wang[4], Karim Lekadir[3], Steffen E Petersen[1], Patricia B Munroe[1]

[1]William Harvey Research Institute, Queen Mary University of London, UK
[2]Aragon Institute of Engineering Research, University of Zaragoza, Spain
[3]Faculty of Mathematics and Computer Science, University of Barcelona, Spain
[4]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

Abstract

Left ventricular hypertrophy (LVH) is an established, independent predictor of cardiovascular morbidity and mortality. Indices derived from the electrocardiogram (ECG) have been used to infer the presence of LVH with limited sensitivity. The aim of this study was to assess the discriminative power of a combination of morphological ECG biomarkers to optimally classify LVH using random forest. As a secondary analysis, we addressed the class imbalance using downsampling. We extracted ECG biomarkers with a known physiological association with LVH from the 12-lead ECG of 38,382 participants in the UK Biobank imaging cohort. LVH was defined based on parameters from cardiac magnetic resonance imaging. In our experiments, the dataset was split into 80% training set for learning and 20% testing set for performance measurement. Additionally, ten-fold cross validation was applied to train the random forest algorithm. Classification of LVH reached 98% accuracy (sensitivity 99%, specificity 4%, F1 score 99%, AUC 0.76) using a combination of 40 ECG biomarkers. Following downsampling of the majority class in the training set, accuracy reached 45% (sensitivity 45%, specificity 89%, F1 score 62%, AUC 0.80). These findings provide support for the ECG to identify LVH using random forest.

1. Introduction

Left ventricular hypertrophy (LVH) is pathologically increased LV mass. LVH is an established independent predictor of cardiovascular morbidity and mortality [1], [2]. Increased LV mass stimulates myocyte hypertrophy, collagen formation and fibroblasts, causing geometric changes and subsequent remodelling of the LV. Consequently, there is a disproportionate increase in myocardial fibrosis which can lead to LV diastolic dysfunction [3][4].

Hypertension is the commonest cause of LVH, accelerating coronary artery atherosclerosis as a substrate for myocardial infarction, scar formation with subsequent LV systolic impairment and malignant arrhythmias. Additionally, there is increased risk of atrial fibrillation, an important marker for congestive heart failure and thromboembolic complications [5].

Cardiac magnetic resonance (CMR) imaging is the gold standard imaging modality in the assessment of LVH. In conditions such as hypertension, the electrocardiogram (ECG) is the first-line diagnostic tool recommended, with the main goal of assessing the presence of LVH [6]. In contrast to CMR, the ECG is an inexpensive screening tool to detect LVH at the bedside, available in both primary and secondary care clinical settings. Indices derived from the ECG have previously shown to detect the presence of LVH but studies consistently demonstrate that ECG based LVH criteria have 30-50% sensitivity [7].

The aim of this study was to assess the discriminative power of a combination of morphological ECG biomarkers to optimally classify LVH and address the class imbalance using downsampling.

2. Methods

2.1. UK Biobank imaging study

The UK Biobank (UKBB) is a prospective population study where demographics, medication history, electronic health records, biomarkers and genomics were collected in half a million participants aged 40-69 years when recruited between 2006 and 2010 from across the United Kingdom. The UKBB imaging study was launched in 2015, with the aim of scanning 20% of the original cohort, that is 100,000 participants. The details of the UKBB CMR protocol have been described elsewhere [8].

A total of 38,382 participants from the UKBB imaging cohort were manually categorised into normal LV and LVH (>70g/m^2 for males and >55g/m^2 for females) based on CMR parameters [9]. The proportion of UKBB participants in each category are shown in table 1.

Table 1. Proportion of UKBB participants in each LV mass category

	Normal LV	LVH
No. of participants	37,731	651
Male (%)	18,130 (48)	343 (53)

2.2. ECG biomarker extraction

We analysed the 10 seconds ECG recording of each of the 38,382 participants using MATLAB to derive the biomarkers with known physiological association with LVH. Only the independent ECG leads (I, II, V1-6) were analysed. In order to perform classification, several representative features were extracted from the signal to compose a feature vector.

Bandpass filtering (1-45Hz) was applied to attenuate baseline wander and high-frequency noise. Following R-wave detection, we applied signal averaging to derive the median ECG waveform of a single beat. To suitably train the models, all ECG biomarkers were normalised with mean and standard deviation to eliminate scale differences during subsequent classification. We also calculated the value of each ECG biomarker across the 8 leads. Several morphological biomarkers were computed directly from the ECG signal including:

- QRS amplitude: computed as the difference between the maximum and minimum points of the QRS complex for each lead
- QRS duration: extracted by deriving QRS onset and offset shown on Figure 1
- Ascending and descending slope of the QRS complex: ascending slope was determined as the upslope from the QRS onset and QRS peak and descending slope as the downslope from QRS peak and QRS offset
- Q wave duration: extracted as the difference between Q wave onset and offset
- Q, R and S wave amplitude: Derivatives were used to detect local minimums in the ECG at tails of QRS ascending slope to determine Q waves and descending slope to determine S wave. The maximum peak determined the R wave
- ST segment deviation: computed as the amplitude from QRS offset to T onset in relation to the isoelectric line
- QT duration: Computed as the difference between Q onset and T wave end
- T wave amplitude: computed as the maximal amplitude following the QRS complex
- T wave duration: extracted by deriving the difference between T wave onset and T wave end. T onset and offset were computed using the tangent method

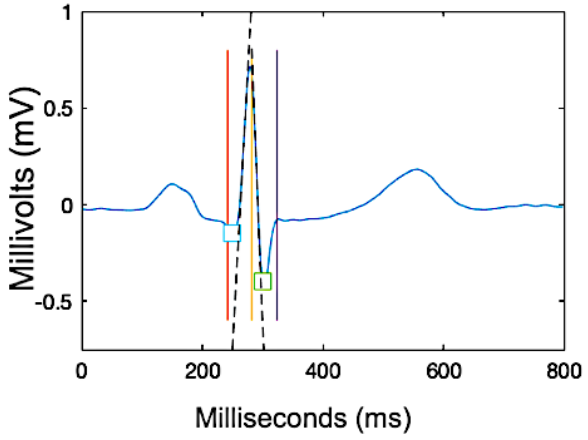

Figure 1. ECG waveform from precordial lead V6 showing QRS onset (orange vertical line), QRS offset (violet vertical line) and QRS ascending and descending slopes (dotted lines)

2.3. Imputation and correlation analysis

ECG biomarkers with less than 10% of missing data were imputed using the MICE package in R. Highly correlated ECG biomarkers were omitted (correlation coefficient threshold of 0.9).

2.4. Machine learning algorithm

In our experiments, the dataset was split into a training set (80%) for learning and a testing set (20%) for performance measurement. Additionally, ten-fold cross validation was applied to train the random forest algorithm.

The feature selection process was performed using Chi-squared to identify the top contributing ECG biomarkers.

A number of key parameters were thoroughly optimised in the training set, including the maximal number of branches as well as the number of features used to split each new node.

Table 2. Ranking of the top 40 ECG biomarkers in the Random Forest algorithm

Top 40 ECG biomarkers included in the Random Forest algorithm			
1. QRS amplitude in V5	11. QRS amplitude in V1	21. ST deviation in V6	31. QT duration in I
2. S amplitude in V3	12. QRS duration in I	22. QRS duration in V2	32. ST deviation in I
3. Global QRS amplitude	13. QRS descending slope in V6	23. QRS duration in V3	33. QT duration in V2
4. S amplitude in V1	14. S amplitude in V4	24. QRS duration in V1	34. QT duration in II
5. QRS amplitude in V4	15. QRS duration in V6	25. QRS amplitude in V1	35. R amplitude in V4
6. Global S amplitude	16. QRS duration in V4	26. QT duration in V6	36. QRS amplitude in II
7. Global QRS duration	17. Global QRS descending slope	27. T amplitude in V6	37. ST deviation in V3
8. R amplitude in V5	18. QRS duration in V5	28. T amplitude in I	38. ST deviation in V2
9. S amplitude in V2	19. Q duration in I	29. Q duration in V5	39. ST deviation in V1
10. Q duration in V6	20. QRS amplitude in V2	30. S amplitude in V5	40. QRS duration in II

As a secondary analysis, in order to address the imbalance in the dataset, downsampling was trialled only in the training set in the majority normal LV group to match the size of the participants in the LVH group (517).

Evaluation of the random forest classifier was implemented in the testing set. The parameters we used to assess classifier performance included: accuracy, sensitivity, specificity, F1 score and Area Under the Curve (AUC).

4. Results and discussion

The combination of the top 40 ranking ECG biomarkers (Table 2) in the random forest classifier were able to discriminate between normal LV and LVH. We found that the top 40 ECG biomarkers provided the optimal model performance. In the imbalanced dataset, accuracy of the classifier was high at 98% but with low specificity at 4%. Downsampling improved specificity to 89% at the cost of reduced accuracy of the classifier. Performance metrics in the test set are presented in Table 3. The ROC curves for the test set in both the imbalanced and downsampled data are shown in Figure 2.

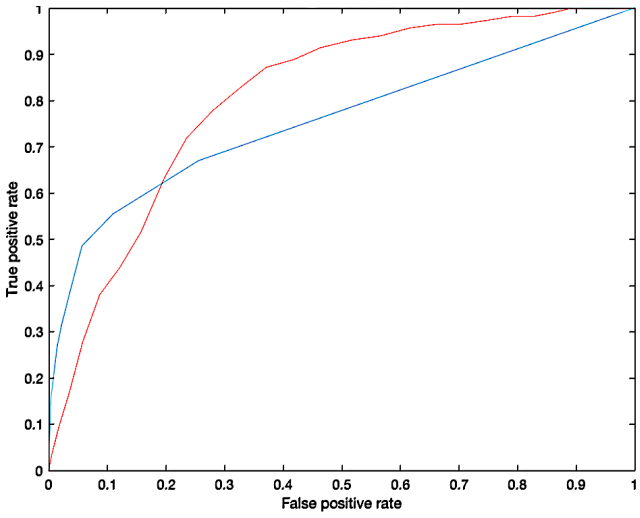

Figure 2. ROC curves for classifying UKBB participants with LVH in the imbalanced test set (blue line, AUC=0.76) and downsampled test set (red line, AUC=0.80)

Table 3. Proportions in the imbalanced and downsampled data with performance measurements in the test set

	Imbalanced data		**Downsampled data**	
	Training	*Test*	*Training*	*Test*
Normal LV	30,189	7,542	517	7,542
LVH	517	134	517	134
Accuracy (%)	98		45	
Sensitivity (%)	99		45	
Specificity (%)	4		89	
F1 score (%)	99		62	

Class imbalance is a common challenge in machine learning, with different techniques proposed to address this issue [10]. Imbalanced datasets degrade the performance of the classifier with the overall accuracy biased to the majority class. The UKBB cohort is a relatively healthy, homogenous population, hence the minority LVH group in the dataset. The random forest classifier may perform better with improved specificity in a different patient population with more cases of LVH.

In our study we selected features with known physiological association with LVH, hence the supervised machine learning approach of using a random forest classifier. In future work, we will be using other classifiers such as support vector machine and exploring more agnostic approaches to ECG feature selection with deep learning for comparison, which has been reported to show promising results [11]. We will also be testing other proposed techniques to address class imbalance and exploring clinical features in addition to ECG biomarkers to improve model performance.

5. Conclusions

In this study we assessed the discriminative power of a combination of extracted morphological ECG biomarkers to optimally classify LVH using random forest and addressed class imbalance in our data using downsampling. Patients with CMR evidence of LVH are at greater risk of cardiovascular events compared with normal LV geometry [12]. Our findings provide support for the ECG as an inexpensive screening tool to identify LVH. This assessment of ECG biomarkers to be able to predict incident cardiovascular outcomes has potential clinical utility.

Acknowledgments

This study was conducted using the UK Biobank resource under access application 2964. We would like to thank all the participants, staff involved with planning, collection and analysis, including core lab analysis of the CMR imaging data. HN is supported by the British Heart Foundation Pat Merriman Clinical Research Training Fellowship (FS/20/22/34640). JR acknowledges funding from the European Union-NextGenerationEU. SEP acknowledges the British Heart Foundation for funding the manual analysis to create a cardiovascular magnetic resonance imaging reference standard for the UK Biobank imaging resource in 5000 CMR scans (www.bhf.org.uk;PG/14/89/31194). SEP and PBM acknowledge support from the National Institute for Health Research (NIHR) Biomedical Research Centre at Barts. S.E.P. and KL have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825903 (euCanSHare project). SEP provides consultancy to and is shareholder of Circle Cardiovascular Imaging Inc., Calgary, Alberta, Canada.

References

[1] D. Levy, "Left ventricular hypertrophy. Epidemiological insights from the Framingham Heart Study," *Drugs*, vol.35 Suppl 5, pp.1–5, 1988.

[2] M. G. Khouri et al, "A 4-Tiered Classification of Left Ventricular Hypertrophy Based on Left Ventricular Geometry: The Dallas Heart Study," *Circ Cardiovasc Imaging*, vol.3, no. 2, pp.164–171, Mar. 2010.

[3] W. S. Aronow, "Hypertension and left ventricular hypertrophy," *Ann Transl Med*, vol.5, no. 15, p.310, Aug. 2017.

[4] F. Sahiti et al, "Left Ventricular Remodeling and Myocardial Work: Results From the Population-Based STAAB Cohort Study," *Front. Cardiovasc. Med.*, vol.8, p.669335, Jun. 2021.

[5] M. Yildiz et al, "Left ventricular hypertrophy and hypertension," *Progress in Cardiovascular Diseases*, vol.63, no. 1, pp.10–21, Jan. 2020.

[6] G. Mancia et al, "2018 ESC/ESH Guidelines for the management of arterial hypertension," p.98.

[7] T. J. Molloy et al, "Electrocardiographic detection of left ventricular hypertrophy by the simple QRS voltage-duration product," *Journal of the American College of Cardiology*, vol.20, no. 5, pp.1180–1186, Nov. 1992.

[8] S. E. Petersen et al, "UK Biobank's cardiovascular magnetic resonance protocol," *J Cardiovasc Magn Reson*, vol.18, p.8, Feb. 2016.

[9] S. E. Petersen et al, "Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort," *J Cardiovasc Magn Reson*, vol.19, no. 1, p.18, Dec. 2017.

[10] S. M. A. Elrahman et al, "A Review of Class Imbalance Problem," Journal of Network and Innovative Computing, 2013.

[11] S. Khurshid et al "Deep Learning to Predict Cardiac Magnetic Resonance–Derived Left Ventricular Mass and Hypertrophy From 12-Lead ECGs," *Circ: Cardiovascular Imaging*, vol.14, no. 6, Jun. 2021.

[12] D. A. Bluemke et al, "The Relationship of Left Ventricular Mass and Geometry to Incident Cardiovascular Events," *Journal of the American College of Cardiology*, vol.52, no. 25, pp.2148–2155, Dec. 2008.

Address for correspondence:

Hafiz Naderi
Department of Clinical Pharmacology, William Harvey Research Institute, Charterhouse Square, London, EC1M 6BQ, United Kingdom
h.naderi@qmul.ac.uk