KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

High-dimensional techniques for ECG analysis

Amalia Villa Gómez

Supervisors:
Prof. dr. ir. S. Van Huffel
Prof. dr. R. Willems
Prof. dr. ir. C. Varon

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

April 2022

High-dimensional techniques for ECG analysis

Amalia Villa Gómez

Examination committee:
Prof. dr. ir. P. Sas, chair
Prof. dr. ir. S. Van Huffel, supervisor
Prof. dr. R. Willems, supervisor
Prof. dr. ir. C. Varon, supervisor
Prof. dr. ir. D. Schreurs
Prof. dr. ir. C. Van Hoof
Prof. dr. ir. A. Bertrand
(KU Leuven)
Prof. dr. ir. C.A. Swenne
(Leiden University Medical Center (LUMC))

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

April 2022

© 2022 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Amalia Villa Gómez , Kasteelpark Arenberg 10 box 2446, 3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Acknowledgments

*"I made it! Or should I say **we** made it?"* This is what I wrote to my supervisors just after sending my thesis to the members of my PhD jury. Now I am even more certain: **we** did it. This beautiful ending of the PhD journey would not have been even imaginable without the help and support of quite a group of people. I hope you all feel this thesis partly yours, you all were essential to make me start, continue and finish this PhD.

First of all, I need to thank Carolina Varon. It is just a fact that I would have not achieved this without your patience, insights, ideas and unconditional support. Probably my favorite moments of the PhD have been our weekly meetings and the awesome feeling I had after them, when I found my research the most beautiful and relevant in the world. Your passion for science, research and, why not say, spectral clustering were contagious and you always gave me the motivation I needed to keep working. I cannot thank you enough for the amount of time and energy you have invested on me, even in your eventful life in the last years. You made me believe I could do this, which you know often was quite a challenge. Gracias infinitas, copromotora linda. For being an example, and always approachable, even to dance flamenco.

I am very happy to include in the **we** responsible of this thesis my supervisors Sabine Van Huffel and Rik Willems. I started this journey with you in my Master thesis, and you believed in me to continue with this PhD. Thank you for the incredible opportunity you have given me, to merge the wide technical knowledge of Sabine's group with the challenges and crucial studies of Rik in the hospital. Thank you for all your time and patience in meetings, for correcting my work, and for your support over these years. I am very proud to be one of the students who worked on your joint collaboration. I have learned a lot from you both and from this journey which would have been impossible without you.

I would also like to thank the members of my supervisory and examination committee: Dominique Schreurs, Chris Van Hoof, Cees A. Swenne and Alexander Bertrand. Your insights and comments have helped us to improve this book; and our discussions made

me think critically about my research and explain it better. I would like to repeatedly thank Alexander Bertrand, for our times in Biomed. Thank you for your gigantic help in the 217 comments on our U2FS algorithm, your patience and for all your comments and questions during the Biomed meetings. Your students are very lucky to work with you. Thanks also to Paul Sas for chairing these defenses.

At some points of this journey, I must say that I could not see the point of what I was doing. At those moments, there were key rescue meetings, which brought inspiration and gave sense to the many hours of computer work. Thank you Sebastian and Bert, for having always the patience to explain me all these cardiology concepts, for having such an open mind for our inventions and for all the cheering up messages. I have really enjoyed our collaborations.

I also need to thank my other big collaborator, the *raja* of patience, algebra and willingness to help. Thank you Abhi for our work together, the online and in-person coffee breaks and the pills of knowledge you always had for us.

My other frequent and inspiring meetings were those with the two great master thesis students I had the honor to supervise. Ben and Tinne, thank you for your hard work and motivation. Your excitement was contagious and you really inspired me for my own research too, as you see in this book.

My first day as a PhD student coincided with the reflection day of Biomed. The group already looked quite promising, and it only got better when every week there was a glass of cava in my hand to celebrate something. Despite the corona times, in which those breaks and plans got more virtual, I feel very lucky to have shared these years with you all. Dries, thank you for being such a cool office mate. I feel we discovered each other a bit late in the PhD, but I am really happy that we finally did. Jo, thank you from the first revisions of my Master thesis, to these times together, sharing fun and struggle. Your energy and smile are contagious, and you always pulled us up. Dorien, thanks for sharing the path, for writing my beknopte samenvatting and for the good times in the office and around. Christos, for the great coffee breaks, dinners, weddings and nice plans. To the researchers + HR of the old and new Biomed, Griet and Cem thank you for your work to make this a fun group! Neetha, thanks for all the nice moments, recipe times, and for revealing to me the secret of the hot chocolate machine. To the Cuban touch in Biomed, that felt like having cousins around, Yissel and Alex thanks for the great times and for bringing that warmth to the office. To the other cousin, the baking star Alex, thanks for the awesome times, the help moving, and the wise tips even from the distance. To the pre-corona Biomed, thanks for the good times of the coffee breaks, with Spyfall, cookies or any other invention, and for the nice conferences and receptions: Ofelie, Kaat, Margot, Jasper, Thomas, the Simons, Andrea, Lieven, Tim, Nick, Jonathan Dan. To the new Biomed, thanks for making it cosy in the office in this new era! Hope you all enjoy as much as we have. And to the essentials: Ida, John, Aldona, Elsy, thank you so much for everything you do for us.

Merging Biomed and my Belgian family, there is a VIP group, the one of my Maricas + copromotora linda, who need a special mention. Mario del puño levantado, thank you so much for being the joy of the office, for your essential presence in the meetings, for your readiness to cheer us up even in your tight political/scientific/sportive agenda. You were my favorite reason to come to the office, no doubt. La compromotora linda, you just deserve a second mention, as friend and family here. Thank you so much for all your care, your jokes. As this thesis, you have been multi-dimensional: in the office, in the fun, in flamenco, as neighbor in Antwerp... I will always be grateful for your *ánimo guerrera* and your *tan boba*. Gracias eternas, y un gracias a Steven por todo el tiempo que os he quitado. Y al hermano por excelencia, John de mis amores, colega de oficina y de la vida belga. Gracias por tu ayuda incondicional, por hacerme siempre sentir genial, por tus cuidados y tus buenas palabras a tiempo. Ha sido un honor y un gusto compartir esta locura contigo. Gracias a Gibrán también, por ofrecernos planazos culturales y por apuntarte siempre a lo que nos inventamos.

And for the weekends, the lunch breaks, the evenings... I need to thank this Belgian family we have created. All our plans, bike rides, BBQ, get-aways, dinners or any amazing time together always made me feel so grateful to have you. Thanks Apek for all the experiences shared, for always finding some time and for all those discussions that made us both feel so understood and in sync. Thanks to the two cool men that you brought to our lives, Antoine and Yuvaan. Plans were always more fun when the dream team was joining. Thank you collega Franci and paixtarísima Adonis for all the great moments and needed distractions. To the first person I met in KU Leuven, Pieri: thank you for always joining the plans, for your uncountable bottles of wine, for showing us Cyprus, teaching us some greek and for making it all the way with me until this moment (and all those to come!). To the brewers, Taranjeet, Ale, Gian, Nicole, thanks for the fun and the good beers. Las cerveceras, escalatas, ciclistas y akelarre, gracias Mar y Mireia por los planazos que nos marcamos. To Saskia, for the good times in CM50, the biking, the rescue food, your cakes and all your care. To the regulars in CM50, Karim, Lucía, Veerle,... thank you for all those fun times around the best dining table in town. And to Elvira, for always joining the good parties and sharing all that fun.

Sometimes the PhD required some creativity, which got reloaded every time I could make some music. Thank you Tom and Jo for letting me be part of your bands and for sharing your amazing music with us. To the Baguetters, Droms and Batterijers, thank you so much for the great times and Tempos. Music also included nice concerts, including international gigs. Thank you to Me, Moon and you for the 3/8s, the jams, and all the music we shared from Herent to Paris.

All this process involved a lot of jumping from one city to another in this tiny but great country. To the Antwerp crew, with or without book, the sushinemas, Ni Shifus and other plans over these years. Thank you Serena, Bernardo, Julia, Nestor, Janan, Alicia, John and Jeremy. Durante este doctorado Bruselas hizo su magia y atrajo a algunas de

mis personas preferidas para cantar a gritos a Silvio, Mercedes Sosa y Rosalía. Mario, Orianne, Taína, Yaumel, mis corazones de melocotón, gracias por las descargas. Y al cubano acoplado, que la vida nos junta y separa, gracias Edu por tus apariciones estelares y los solos de olla exprés.

Sin duda he cogido fuerzas cada vez que iba a casa, o la casa venía a mi. Muchísimas gracias a todas las personas increíbles que habéis pasado por mi casa en este tiempo, y a quienes siempre habéis estado deseando recibirme cuando volvía a España. A los abuelos y al tito Edu, por preguntarme lo primero siempre cuánto tiempo me quedaba y cuándo volvía, porque salir a comer con vosotros y los abrazos que me dáis me dan la vida. A las titas Merce y Tamara, por venir, por cuidarme, por estar orgullosas siempre y enseñarme tanto. A los Gómez, por las Carabas, Herraduras y por ser increíbles. Os quiero a todos muchísimo. A los de siempre: Bea, Julio, gracias por cada vez que habéis venido y las que nos habéis recibido. A Héctor por los chistes siempre a tiempo y por el grupo de madres de 4ºA. A Iván, por los conciertos y las capturas de piñatas. A Carmela, Carlos, Adri, Vero, Daniela y Álvaro, por entender mi ritmo desquiciante y siempre alegrarse de verme. Trataré de ir más despacio a partir de ahora. A Elo y Marta, que vinieron y vendrán, por cuidarnos y redescubrinos. A Maxime, por enseñarnos Bruselas, el Vercors y la sierra de Madrid, y por las ovaciones de muchedumbres. A Yenni, por los mil enredos, las buenas fiestas y los siempre buenos planes. A Faus, por las pelis, por los paseos y por seguir ahí. A Margui y Paula, for the awesome whatsapp messages and the craziest night of these PhD years. A Belén, porque me ayudaste a re-encauzar esta tesis y me enseñaste muchísimo.

Hay por ahí unas ingobernables, que son inspiración, compañeras de la lucha, titanas e inmensamente cuidadosas, sin las que no hubiera hecho esto. Siempre hacéis que me sienta acompañada, fuerte y casi tan especial como vosotras. Luci, gracias por venir siempre a la locura que sea, por tu música y por enseñarme a disfrutar. Tere, por los podcasts, las cervezas que no son cerveza, y por inspirarme tantísimo. Lara, porque tu compañía es esencial en ese estar perdidas, y cada camino hay que hacerlo juntas, por si nunca nos encontramos. Paula, por tus ojos que siempre me entienden, por recibirme en todos los lugares en los que has vivido, por las sorpresas, nuestro historión y por seguir haciendo todo mucho más, mucho mejor.

In the tight, tight *we*, almost writing part of this thesis, or at least making me sit down to do it, listening to my endless overthinking, preparing a million parties, surprises, meals. I had the best possible family in Belgium, the JAKES team. Kranti, my soulmate, I really couldn't have done this without you. Thank you for all your wisdom, for being always ready to listen even when you could not understand me, and for being so awesome. If I need to think of a role model, it may be you. Thank you Shradha, for all your care, your chais, the games and your sweetness. Jorge, por venir conmigo, por empezar y terminar esto juntos, como no podía ser de otra manera. Por todo lo que me cuidas y todo lo que compartes conmigo, por lo bien que lo hemos pasado. Estoy orgullosísima de dónde estamos.

Esteban, sólo puedo darte el *gracias por todo* más literal de la historia. Gracias por traerme a Leuven, por motivarme siempre a que consiga lo que me proponga, por cuidarme tantísimo, por quererme tan increíblemente y por tu infinita paciencia conmigo. Ahora nos toca seguir disfrutando y continuar con la siguiente temporada.

Y a los últimos responsables de esta experiencia, al equipo perfecto. Gracias Ana por inventar buenos planes, por hacerme siempre reír, por venir y acompañarme tantísimo. Muchísimas gracias por cuidar a todos mientras yo no estaba, por hacer los mejores regalos y por ser así de valiente con todo lo que haces. Gracias José Luis, por cuidarla a ella y por venir hasta aquí. Y a mamá y papá, porque sin vosotros esto hubiera sido imposible. Por todas las oportunidades que nos habéis dado, y por educarnos así de libres, fuertes y darnos el mejor ejemplo. Por vuestro amor y vuestro apoyo, que me llegan siempre, este doctorado es vuestro. Os quiero muchísimo.

Abstract

The electrocardiogram, or ECG, is a key diagnostic tool in clinical practice, due to its simple and non-invasive way of recording the electrical activity of the heart. In clinical examinations, the ECG is mostly captured using 12 leads to obtain a multi-view graphical representation of the cardiac cycle. These signals (i.e., time series) are recorded from different angles, to characterize the electrical activity of the different regions of the heart.

ECG analysis allows identifying abnormalities that can be associated to different cardiac diseases. The presence of some of these abnormalities can be linked to a high risk of arrhythmia and mortality. Therefore, it is crucial to detect these aberrant patterns in time to ensure treatment and prevent the occurrence of life threatening events. One of the main challenges in cardiology is the prediction of ventricular arrhythmia. For that aim, several ECG biomarkers have been proposed in literature, and algorithms have been developed for their automatic detection. However, many of these biomarkers do not make use of the complete 12-lead information, by either limiting their analysis to one of the leads, or deriving 12 independent features that need to be jointly interpreted by clinicians. One of the challenges of using 12-lead ECG is that, while the 12-leads provide different views with complementary information to characterize the entire cardiac cycle, they also contain redundant information. Therefore, robust algorithms are required for the analysis of these multi-lead signals, that can discriminate risk biomarkers from other irrelevant information, while maximally exploiting the content of the 12-leads. Some of these approaches imply the use of dimensionality reduction techniques to translate the information contained in 12-lead ECG into simpler data structures. Additionally, powerful models for arrhythmia or mortality risk prediction are multi-variate, i.e. they are a combination of more than one biomarker or feature that contribute to predict the outcome. These models can also benefit from dimensionality reduction techniques to decrease the number of features, thereby increasing their interpretability and accuracy.

This thesis presents advanced algorithms for the analysis of multi-dimensional data, both for feature-based approaches and for time series analysis. The main motivation

in the development of these algorithms was to create reliable tools usable in clinical practice, which can support further diagnosis. In particular, these methods are defined to be used on ECG analysis, for the identification of arrhythmia risk predictors.

The first part of this thesis proposes two robust algorithms for the extraction of ECG features for arrhythmia risk prediction. These algorithms include strategies to reduce the impact of noise in the analysis of the signal and the influence of ambiguous annotations. First, a heartbeat classification algorithm for single-lead ECG was proposed. It achieved results comparable to the state-of-the-art, while facing the same challenges related to the accurate detection of under-represented classes. This triggered an exhaustive analysis of the database and the protocol used for training and evaluating heartbeat classification methods, highlighting their limitations. The second algorithm proposed for ECG analysis focuses on an improved quantification of QRS fragmentation (fQRS). This method was trained and evaluated on multi-center data, exploring the impact of including different fQRS definitions in the training process, hence tackling the problem of inter-observer variability. Results indicate that the combination of two different fQRS definitions achieved comparable results to those trained and evaluated on a single definition. Therefore, this study confirms the relevance of using multi-center data in the development of these algorithms, to make them more flexible and applicable to clinical practice.

The ECG analysis algorithms mentioned above, as well as most of the state-of-the-art methods, analyze each ECG-lead independently. Therefore, the second part of the thesis aims to maximally exploit the information contained in multi-lead ECG analysis by defining a new framework for ECG dimensionality reduction based on Laplacian Eigenmaps. This data-driven approach emphasizes the irregularities in the data, and it was evaluated for the characterization of structural and arrhythmogenic pathologies from 12-lead ECGs. This framework was compared against the vectorcardiogram, which is a 3-lead signal that can be derived from the 12-lead ECG. Results suggest that the LE framework is more robust to outliers than the approach based on VCG. This confirms the advantages of data-driven approaches for 12-lead ECG dimensionality reduction, setting the path for more efficient and exhaustive analysis of multi-lead signals. The obtained reduced-dimensionality signals preserve and enhance abnormalities, which are often missed both by visual and current automated analysis.

The last part of this thesis focuses on a feature-based approach and it presents a novel unsupervised non-parametric feature selector: the U2FS algorithm. This user-friendly method based on the utility metric was evaluated using simulated and benchmark databases and it was proven to obtain results in line with the state-of-the-art, while requiring less computation time.

Beknopte samenvatting

Het elektrocardiogram, of EKG , is een belangrijk diagnostisch hulpmiddel in de klinische praktijk, vanwege de eenvoudige en niet-invasieve manier om de elektrische activiteit van het hart vast te leggen. Bij klinische onderzoeken wordt het EKG meestal opgenomen met 12 afleidingen om een grafische weergave van de hartcyclus in meerdere weergaven te verkrijgen. Deze signalen (d.w.z. tijdreeksen) worden vanuit verschillende hoeken geregistreerd om de elektrische activiteit van de verschillende gebieden van het hart te karakteriseren.

EKG-analyse maakt het mogelijk afwijkingen te identificeren die kunnen worden geassocieerd met verschillende hartaandoeningen. De aanwezigheid van sommige afwijkingen kan in verband worden gebracht met een hoog risico op aritmie en sterfte. Daarom is het van cruciaal belang om deze afwijkende patronen op tijd te detecteren om dit te kunnen behandelen en het optreden van levensbedreigende gebeurtenissen te voorkomen. Een van de belangrijkste uitdagingen in de cardiologie is het voorspellen van ventriculaire aritmie. Hiervoor zijn in de literatuur verschillende EKG-biomarkers voorgesteld en zijn algoritmen ontwikkeld voor de geautomatiseerde detectie. Veel van deze biomarkers maken echter geen gebruik van de compleet informatie van 12 afleidingen. Ze beperken hun analyse ofwel tot één van de afleidingen, ofwel door 12 onafhankelijke kenmerken af te leiden die door clinici als een geheel moeten worden geïnterpreteerd. Hoewel de 12 afleidingen verschillende weergaven bieden met aanvullende informatie om de hele hartcyclus te karakteriseren, bevatten ze echter ook overbodige informatie. Daarom zijn robuuste algoritmen vereist voor de analyse van deze multi-afleidingen signalen, die risicobiomarkers kunnen onderscheiden van andere irrelevante informatie en de inhoud van de 12-afleidingen maximaal benutten. Sommige van deze benaderingen impliceren het gebruik van dimensionaliteitsreductietechnieken om de informatie in 12-afleidingen EKG te vertalen naar eenvoudigere gegevensstructuren. Bovendien zijn krachtige modellen voor het voorspellen van aritmie of sterfterisico multivariaat, d.w.z. ze zijn een combinatie van meer dan één biomarker of kenmerk die bijdragen aan het voorspellen van de uitkomst. Deze modellen kunnen ook profiteren van dimensionaliteitsreductietechnieken om het aantal kenmerken te verminderen, waardoor

hun interpreteerbaarheid en nauwkeurigheid toenemen.

Dit proefschrift presenteert geavanceerde algoritmen voor de analyse van multidimensionale gegevens, zowel voor op kenmerken gebaseerde benaderingen als voor tijdreeksanalyse. Het belangrijkste doel van deze algoritmen is om te dienen als betrouwbare hulpmiddelen in de klinische praktijk, die verdere diagnose kunnen ondersteunen. Deze methoden zijn met name gedefinieerd om te worden gebruikt bij EKG-analyse, voor de identificatie van voorspellers van aritmierisico's.

Het eerste deel van dit proefschrift presenteert twee robuuste algoritmen voor de extractie van EKG-kenmerken voor de voorspelling van het risico op aritmie. Deze algoritmen omvatten strategieën om de impact van ruis bij de analyse van het signaal en de invloed van dubbelzinnige annotaties te verminderen. Ten eerste werd een classificatie-algoritme voor de hartslag voor EKG met één enkele afleiding voorgesteld. Het behaalde resultaten die vergelijkbaar zijn met de state-of-the-art, terwijl ze voor dezelfde uitdagingen stonden met betrekking tot de nauwkeurige detectie van ondervertegenwoordigde aritmieklassen. Dit leidde tot een grondige analyse van de database en het protocol dat wordt gebruikt voor het trainen en evalueren van classificatiemethoden voor hartslagen, waarbij hun beperkingen worden benadrukt. Het tweede voorgestelde algoritme voor EKG-analyse richt zich op een verbeterde kwantificering van QRS-fragmentatie (fQRS). Deze methode is getraind en geëvalueerd op data uit meerdere centra, waarbij de impact is onderzocht van het opnemen van verschillende fQRS-definities in het trainingsproces, waardoor het probleem van variabiliteit tussen waarnemers wordt aangepakt. De resultaten geven aan dat de combinatie van twee verschillende fQRS-definities vergelijkbare resultaten opleverde als degenen die waren getraind en geëvalueerd op een enkele database. Daarom maakt deze trainingsstrategie het algoritme flexibeler en toepasbaar in de klinische praktijk.

De hierboven genoemde EKG-analyse-algoritmen, evenals de meeste geavanceerde methoden, analyseren elke EKG-afleiding onafhankelijk. Daarom is het tweede deel van het proefschrift gericht op het maximaal benutten van de informatie in multi-afleidingen EKG-analyse door een nieuw raamwerk te definiëren voor EKG-dimensionaliteitsreductie op basis van Laplace Eigenmaps. Deze gegevensgestuurde benadering benadrukt de onregelmatigheden in de gegevens en werd geëvalueerd voor de karakterisering van structurele en aritmogene pathologieën van 12-afleidingen EKG's. Dit raamwerk werd vergeleken met het vectorcardiogram, een 3-afleidingen signaal dat kan worden afgeleid uit het 12-afleidingen EKG. De resultaten suggereren dat het LE-raamwerk robuuster is voor uitschieters dan de benadering op basis van VCG. Dit bevestigt de voordelen van gegevensgestuurde benaderingen voor het verminderen van de dimensionaliteit van 12-afleidingen EKG's, waarmee de weg wordt vrijgemaakt voor een efficiëntere en uitgebreidere analyse van multi-afleidingen signalen. De verkregen signalen met verminderde dimensionaliteit behouden en versterken afwijkingen, die vaak worden gemist door zowel visuele als huidige geautomatiseerde analyse.

Het laaste deel van dit proefschrift richt zich op een kenmerk gebaseerde benadering en presenteert een nieuwe niet-gesuperviseerde niet-parametrische feature selector : het U2FS-algoritme. Deze gebruiksvriendelijke methode, gebaseerd op de utiliteitsmetriek, werd geëvalueerd met behulp van gesimuleerde en benchmarkdatabases . Dit leverde resultaten op die overeenkwamen met de modernste technologiën , terwijl er minder rekentijd nodig was.

Abbreviations

Abbreviation	Meaning
AAMI	Association for Advancement of Medical Instrumentation
ACC	Accuracy
AUC	Area Under the Curve
AV	Atrioventricular
CinC	Computing in Cardiology
DCM	Dilated Cardiomyopathy
DS1	Training-set MIT-BIH arrhythmia database
DS2	Evaluation-set MIT-BIH arrhythmia database
DTW	Dynamic Time Warping
ECG	Electrocardiogram
EEG	Electroencephalogram
ES	Electrical Storm
ESP	Extended Sammon Projection
EU-CERT-ICD	EUropean Comparative Effectiveness Research to Assess the Use of Primary ProphylacTic Implantable Cardioverter Defibrillators
F	Fusion beat
fQRS	QRS fragmentation
HOS	Higher Order Statistics
HRV	Heart Rate Variability
HRT	Heart Rate Turbulence
ICA	Independent Component Analysis
ICD	Implantable Cardioverter-Defibrillator

IDT	Inverse Dower Transform
IHD	Ischemic Heart Disease
KNN	K-nearest neighbours
LARS	Least-Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
LBBB	Left Bundle Branch Block
LE	Laplacian Eigenmaps
LS	Least Squares
LSE	Least Squares Error
LS-SVM	Least-Squares Support Vector Machines
MCFS	Multi-cluster Feature Selection
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
N	Normal beat
NDFS	Nonnegative Discriminative Feature Selection
PCA	Principal Component Analysis
PDF	Probability Density Function
PPV	Positive Predictive Value
PR	Precision-Recall
PRSA	Phase-Rectified Signal Averaging
Q	Unknown beat
RBBB	Right Bundle Branch Block
RBF	Radial Basis Function
RJGSC	Robust Joint Graph Sparse Coding
ROC	Receiver Operator Characteristic
SA	Sinoatrial
Sens	Sensitivity
SCD	Sudden Cardiac Death
SCP	Standard Communications Protocol
SD	Standard Deviation
SHAP	SHapley Additive ExPLanations
Spec	Specificity
SSFS	Sparsity-inducing Feature Selectors
SVD	Singular Value Decomposition

SVEB	Supraventricular Ectopic Beat
SVM	Support Vector Machines
UZL	University Hospitals Leuven
VCG	Vectorcardiogram
VEB	Ventricular Ectopic Beat
VF	Ventricular Fibrillation
VMD	Variational Mode Decomposition
VT	Ventricular Tachycardia
Vx	Chest electrodes V1 to V6

Contents

Abstract vii

Beknopte samenvatting ix

Contents xvii

List of Figures xxi

List of Tables xxv

1 Introduction 1
- 1.1 Research motivation . . . 1
- 1.2 Background . . . 4
  - 1.2.1 The heart . . . 4
  - 1.2.2 The cardiac electric cycle and the ECG . . . 5
  - 1.2.3 Aberrant patterns in the ECG: abnormal heartbeats and arrhythmia . . . 7
- 1.3 ECG biomarkers for ventricular arrhythmia prediction: previous work 10
- 1.4 Research objectives . . . 12
- 1.5 Chapter overview and personal contributions . . . 13
- 1.6 Collaborations . . . 17

2 Heartbeat classification in single-lead ECG 19
- 2.1 Introduction . . . 20
- 2.2 Methods . . . 21
  - 2.2.1 Data . . . 21
  - 2.2.2 Pre-processing and heartbeat segmentation . . . 22
  - 2.2.3 Feature extraction . . . 23
  - 2.2.4 Feature selection . . . 25
  - 2.2.5 Classifier and performance metrics . . . 25
- 2.3 Results and discussion . . . 25

2.3.1 Selected features . . . 25
2.3.2 Classification results . . . 26
2.3.3 Feature analysis . . . 27
2.3.4 Alternative results . . . 28
2.4 Conclusion . . . 29

3 Fragmented QRS Dynamics Towards Electrical Storm in ICD Patients 31
3.1 Introduction . . . 32
3.2 Methods . . . 32
3.2.1 Preprocessing and Segmentation . . . 33
3.2.2 Feature extraction . . . 34
3.2.3 Automatic quantification of QRS fragmentation . . . 35
3.2.4 Statistical Analysis . . . 36
3.3 Results . . . 36
3.3.1 Differences between ES and control patients in the last visit . 36
3.3.2 Differences between the evolution of ES and control patients . 37
3.4 Discussion . . . 37
3.5 Conclusion . . . 39

4 A machine learning algorithm for electrocardiographic fQRS quantification validated on multi-center data 41
4.1 Introduction . . . 42
4.2 Methods . . . 43
4.2.1 Data . . . 43
4.2.2 Pre-processing . . . 46
4.2.3 QRS segmentation . . . 46
4.2.4 Feature extraction . . . 47
4.2.5 Classifier . . . 50
4.2.6 Experiments . . . 51
4.2.7 Performance metrics . . . 52
4.2.8 Analysis of feature importance . . . 54
4.3 Results . . . 54
4.3.1 Experiment 1: The impact of the new segmentation algorithm 54
4.3.2 Experiment 2: The effect of different fQRS annotation criteria 57
4.3.3 Experiment 3: Using the fQRS algorithm in irregular AF recordings . . . 58
4.3.4 Analysis of feature importance . . . 59
4.4 Discussion . . . 59
4.5 Conclusion . . . 64

5 Manifold learning for 12-lead ECG dimensionality reduction: A unified framework based on Laplacian Eigenmaps. 65
5.1 Introduction . . . 66

5.2 Methods . . . 67
5.2.1 Data . . . 67
5.2.2 Preprocessing . . . 69
5.2.3 Definition of the reference LE space . . . 70
5.2.4 Explicit mapping . . . 73
5.2.5 Feature extraction . . . 73
5.2.6 Experiments . . . 76
5.2.7 Statistical tests . . . 77
5.3 Results . . . 78
5.3.1 Capturing characteristics of normal ECGs . . . 78
5.3.2 Identifying cardiovascular diseases: differences between IHD and DCM . . . 79
5.3.3 LE biomarkers for arrhythmia prediction: electrical storm . . . 80
5.4 Discussion . . . 81
5.5 Conclusion . . . 85

6 Vectorcardiogram analysis for patient characterization 87
6.1 Introduction . . . 88
6.2 State-of-the-art: VCG for IHD, DCM and arrhythmia prediction . . . 90
6.2.1 VCG in IHD and DCM patients . . . 90
6.2.2 VCG and arrhythmia prediction . . . 91
6.3 Methods . . . 91
6.3.1 Preprocessing . . . 92
6.3.2 Feature extraction . . . 93
6.3.3 Experiments . . . 100
6.3.4 Statistical Tests . . . 100
6.4 Results . . . 101
6.4.1 Capturing characteristics of normal ECG . . . 101
6.4.2 Identifying cardiovascular diseases: differences between IHD and DCM . . . 102
6.4.3 Arrhythmia prediction: Electrical storm . . . 105
6.5 Discussion . . . 106
6.6 Conclusion . . . 109

7 Manifold learning for unsupervised feature selection: the U2FS algorithm 111
7.1 Introduction . . . 112
7.2 Related Work . . . 114
7.3 Methods . . . 115
7.3.1 Manifold learning considering non-linearities . . . 115
7.3.2 Kernel parameter approximation for high-dimensional data . . . 117
7.3.3 Utility metric for feature subset selection . . . 120
7.4 Results . . . 124

7.4.1 Simulations . . . 126
7.4.2 Benchmark datasets . . . 128
7.5 Discussion . . . 132
7.6 Conclusion . . . 135

8 Conclusions and future work **137**
8.1 Conclusions . . . 137
8.1.1 Robust algorithms for the extraction of arrhythmia risk predictors from the ECG . . . 138
8.1.2 Framework for dimensionality reduction of 12-lead ECG signals 139
8.1.3 Unsupervised feature selector based on manifold learning . . 139
8.2 Future work . . . 140
8.2.1 Robust algorithms for the extraction of arrhythmia risk predictors from the ECG . . . 140
8.2.2 Framework for dimensionality reduction of 12-lead ECG signals 141
8.2.3 Unsupervised feature selector based on manifold learning . . 142

A Appendix **143**
A.1 Supplementary material for Chapter 5 . . . 143
A.1.1 Table for results of different kernels, Experiment 1 . . . 143

Bibliography **145**

Curriculum Vitae **167**

List of Figures

1.1 Schematic image of the anatomy of the heart. 5
1.2 Steps of the cardiac cycle and relation with the ECG waves. 6
1.3 Standard ECG lead configurations. 8
1.4 Example of a 12-lead ECG signal. 9
1.5 Examples of non-fragmented (left) and fragmented QRS complexes. . 11
1.6 Schematic view of this thesis. 14

2.1 Examples of heartbeat classes and their corresponding variational modes. 24
2.2 Classification error per number of features removed. 26
2.3 RR distance value distribution for the SVEB class. 28

3.1 Global fragmentation scores for the last visit of ES patients and control patients. 37
3.2 Average fragmentation score along the visits, from 3 years before to the time around electrical storm. 38

4.1 Block diagram of the fQRS scoring algorithm per lead. 43
4.2 Example QRS segmentation for multi-lead signal with irregular heartbeats. 48
4.3 Examples of the VMD decomposition of non-fragmented and fragmented QRS complexes. 49
4.4 Examples of the PRSA procedure for non-fragmented and fragmented QRS complexes. 50
4.5 Block diagram describing the training strategies to evaluate the effect of multiple fQRS annotation criteria. 53
4.6 ROC curves for the proposed and reference approaches applied to test1 signals of UZL dataset, exclusively including signals without inter-observer variability. 55

4.7 Relation between the output scores of the proposed and reference approaches (in black and grey) with the labels of test2 UZL signals, ranging from 0 to 5. 56
4.8 ROC and PR curves for test1 UZL and EU-CERT for the 3 training strategies considered. 58
4.9 SHAP analysis of UZL and EU-CERT test signals for the combined model. 60
4.10 Examples of missclassified QRS complexes. 63

5.1 Preprocessing steps of the ECG signals prior to applying the LE algorithm. 69
5.2 Structure of the data used to build the matrix required for the LE algorithm. 71
5.3 Trajectories of the ECGs used to build the reference space. 72
5.4 Examples of reference heartbeats explicitly mapped on the LE space for ES and non-ES patients. 74
5.5 Violin plots for d_{DTW}, r_{V2} and σ_{V2} for the 10 groups of normal ECGs. 78
5.6 Violin plots for the features presenting significant differences for experiments 2 and 3. 80
5.7 Relation between QRS duration and the features for the cases of higher correlation. 81
5.8 Example of 3 QRS complexes in the LE space and their corresponding ECGs. 84

6.1 Electrode positions for Frank VCG recording. 88
6.2 Example of VCG loops in 3D space. 89
6.3 Preprocessing steps prior to the VCG feature extraction. 92
6.4 Example segmentation. 93
6.5 Graphic representation of the QRS-T angle. 95
6.6 Example of a QRS loop with a normal pattern. 97
6.7 Steps to extract the Feret Ratio of two example loops. 98
6.8 Scheme of the quadrants related to the anatomical planes. 99
6.9 Violin plots of the normal ECGs for QRS-T angle, FQ1 and QRS-axis. 101
6.10 Correlation of QRS duration and speed SD. 103
6.11 Distribution of the QRS-loops areas on the different quadrants. 104
6.12 Violin plots for the Feret ratio in the frontal plane (FRF) and the QRS-T angle. 105
6.13 Curvature values for the 10 H_i groups and for the patients. 106

7.1 Weight system for relevance estimation. 119
7.2 Toy examples used for simulations. 126
7.3 Accuracy results for the benchmark databases, for selecting from 10 to 80% of the original number of features. 130

7.4 Computation time for extracting from 10 to 80% of the original number of features for each of the benchmark databases. 131

List of Tables

2.1 Heartbeat classes and relation with MIT-BIH Arrhythmia dataset definitions. 22
2.2 Description of the MIT-BIH arrhythmia database. 22
2.3 List of features for heartbeat classification. 23
2.4 Selected features in order of relevance 26
2.5 Results on MIT-BIH (DS2) 27
2.6 Classification results for different training-test sets 29

3.1 Number of subjects present in each visit 33

4.1 Patient characterization for UZL and EU-CERT datasets. 44
4.2 Percentage of signals per label in the UZL and EU-CERT datasets . . 45
4.3 List of features extracted from each lead. 50
4.4 Results of the proposed and reference approach on test1 of UZL including only signals with agreed labels between annotators. 55
4.5 Summary of the results of the proposed and reference approach in test2. 56
4.6 Results of the 3 training strategies to evaluate the effect of multiple fQRS annotation criteria. 57
4.7 Results for the combined model in test1 UZL and EU-CERT broken down by QRS duration. 58
4.8 Results of the combined model obtained from Experiment 2 in AF signals of both UZL and EU-CERT. 59

5.1 Baseline demographics of the databases for LE analysis. 68
5.2 Features derived from the LE trajectories 75
5.3 Results for H_1, IHD, DCM, ES and non-ES. 79

6.1 Features derived from the VCG loops 94

7.1 Methods compared in the experiments 125
7.2 Description of the toy example datasets. 127

7.3 Results feature selection for toy examples. 128
7.4 Description of the benchmark databases. 128
7.5 Comparison of classification accuracy (ACC) with the state-of-the-art for PCMAC and BASEHOCK datasets. 131

A.1 Results for Experiment 1 for each of the 3 kernels considered in the SVM classifier. 144

Chapter 1

Introduction

The human heart has hidden treasures,
in secret kept, in silence sealed.
The thoughts, the hopes, the dreams, the pleasures,
whose charms were broken if revealed.

Charlotte Brontë

1.1 Research motivation

The heart is the center of the circulatory system, responsible of pumping blood to the entire body to distribute oxygen and nutrients, and to remove carbon dioxide and other residuals. It has indeed some hidden secrets, but the keys of its functioning have been thoroughly studied for centuries. Far from breaking any charm, cardiac research can only improve the quality of life of patients and reduce the burden of cardiovascular diseases.

Electrocardiography (ECG) allows to visualize the electrical activity of the heart and assess its functioning using a set of electrodes placed on the chest. It is a key diagnostic tool in clinical practice, due to its simple, non-invasive way of recording, that allows capturing a great amount of relevant information about the cardiac cycle. This signal not only captures information about the functioning of the heart, but also about other processes like respiration or the activity of the autonomous nervous system. Therefore, the use of ECG in clinical practice is very extended; and its interpretation has been studied and improved for decades.

Despite the short recording time of the ECG during clinical examinations, this signal has a high diagnostic power. This time window, which is usually 10 seconds, may not be sufficiently long to capture aberrant patterns in the heart functioning of a patient. For this reason, many solutions have explored capturing longer signals using more comfortable set-ups, moving these recordings out of the hospital environment. However, these longer signals can be strongly affected by artefacts and noise, since the patient is not in a controlled environment, neither resting during the recording. Additionally, the analysis of these signals implies extra challenges, due to the lack of complementary information describing the context of these recordings, i.e., the activities that were performed while recording, or the presence of interference sources. Therefore, ECGs recorded during hospital visits are still crucial in the diagnosis of cardiac disease.

Clinical ECG examinations often use the 12-lead modality, allowing the acquisition of the electrical activity of the heart from multiple points of view by using electrodes placed on different locations of the chest. All these electrodes record the cardiac cycle from different positions, capturing details about the different regions of the heart. This multi-view approach enables a detailed description of the cardiac activity, but the increased number of leads presents some challenges both for visual and automated analysis. First, there is redundant information, since all leads are observations of the same underlying process. Part of these redundancies stems from the 12-lead recording set-up, in which only 8 out of the 12 leads are independent from each other, and the other 4 are linear combinations of these 8. Second, each of the independent leads may be affected differently by electrode movement or other sources of artefacts. This contamination hampers the analysis of the signals, since it can mask patterns in some relevant leads, reducing the quality of the data. Hence, the interpretation of these signals requires identifying relevant patterns by merging the complementary information present in the leads, discarding redundancies for simpler analysis and excluding artefacts.

In terms of signal processing, a 12-lead ECG signal can be considered a multi-dimensional problem, where each lead can be interpreted as one dimension. This means that every data sample of the signal is described by 12 different values linked to each of the leads, resulting in a 12-dimensional space. These dimensions should not be understood as geometric, but as an alternative representation of the data. Imagine a room, where every point in the space can be described by its X, Y, and Z position. These dimensions inform about the location of this point. However, each of these points could be characterized by other properties as well, such as the temperature at that point, the humidity, the intensity of the wind, its direction,... This set of descriptors result in a multi-dimensional space, where every point is described by a vector of multiple characteristics. Similarly, every data point of an 12-lead ECG is described by a 12-dimensional vector. Among these dimensions, not all of them would be relevant for every analysis, some could be redundant like in the case of the 4 linearly dependent leads of 12-lead ECG, and some may be randomly affected by noise. Additionally,

high-dimensional data implies interpretability challenges, which get more complicated with the increasing number of dimensions. Hence, multi-dimensional data analysis, both visual or automated, often benefits from dimensionality reduction methods. These approaches compact the information contained in high-dimensions, in this case in 12-lead ECG, into a lower number of dimensions. This step allows using simpler methods to analyze the data while preserving its relevant content.

Different algorithms have been proposed to reduce the dimensionality of the 12-lead ECG, with the aim of removing redundant information with a limited information loss [96, 131]. For instance, several approaches allow deriving the vectorcardiogram (VCG) from 12-lead ECG, obtaining a set of 3 orthogonal leads that represent the instantaneous strength and direction of the cardiac dipole [112, 188]. Most of these methods apply a fixed mathematical transformation that combines the information in 12-lead ECG without considering the specific characteristics of the signals [49, 95, 188]. More recent works proposed patient- or population-specific transformations, which were proven to lead to more accurate VCG representations and adapt to the signal characteristics [40, 114, 189].

Data-driven dimensionality reduction methods for time-series offer a trade-off between generic and patient-specific transformations. These methods make use of mathematical approaches for signal decomposition to reduce dimensionality based on the information contained in the signal. Some examples are the use of Singular Value Decomposition (SVD) to remove noise and obtain a 3-dimensional signal [1], or Laplacian Eigenmaps (LE) for the exploration of electrograms in ischemia [63]. However, these methods do not offer a common framework where signals of different patients can be compared, due to these signal-specific transformations.

Dimensionality reduction approaches can also be applied to other high-dimensional data, such as medication, demographics and other features extracted from biomedical signals or clinical records. In this type of data, one patient is represented by a long set of variables, where each variable represents one dimension (similarly to the example previously mentioned about the points in a room). For instance, in [99] a model for prediction of imminent ventricular arrhythmia is proposed based on several features extracted from the ECG and respiration signal, and in [76] ECG-derived features are combined with vital signs and demographic data for a risk stratification of major adverse cardiac events. While the use of several variables can enable a holistic analysis, this often comes with data redundancy or extra information which could be unrelated to the problem under investigation. Therefore, this type of data can also benefit from dimensionality reduction approaches, where the most relevant dimensions (i.e., features) can be identified while removing any redundancy. Multiple algorithms exist to achieve this goal but often, they are computationally expensive [133, 207] or they are formulated in a supervised manner [45, 143]. The latter relies on accurate ground truth, which is often difficult to get due to the extensive amount of time that is required to annotate medical data. In addition, visual inspection of the data often misses crucial

patterns that might lead to the identification of diseases [72, 175]. Therefore, data-driven approaches become more and more relevant to assist clinical diagnosis and broaden the potential of current automated ECG analysis [159, 205].

This thesis presents advanced algorithms for the analysis of multi-dimensional data, both for time series and feature-based approaches. The main goal of these algorithms is to serve as reliable tools in clinical practice, providing interpretable results to support diagnosis for the physicians. Some of these algorithms have been made publicly available and have a wide range of impact, since they have been formulated in a general way so that they also can be applied to data different from ECG. Most of the studies in this thesis focus on the application of these methods to ECG analysis, with the goal of unveiling some of the secrets yet kept in the heart, especially for the detection of abnormalities. The technical background of the algorithms is provided in their corresponding chapters, but their clinical motivation is presented in the next sections.

1.2 Background

The studies conducted in this thesis are mostly related to the analysis of the electrical activity of the heart, belonging to the field of cardiology. This section provides background knowledge in this field, to set the basis for the different cardiac applications considered in this thesis. The explanations were mostly taken from [31, 51, 142].

1.2.1 The heart

The heart is the central organ of the circulatory system. It pumps blood containing nutrients and oxygen to the entire body, and it sends waste and carbon dioxide to the kidneys and the lungs to secrete them. It must constantly work to irrigate the entire body and keep cells alive.

It is formed by four chambers, two superior called left and right atria, and two inferior called left and right ventricles. An illustration of the heart can be seen in Figure 1.1. Non-conductive tissue separates these chambers and keeps them electrically isolated from each other. The right side of the heart pumps blood to the lungs for its oxygenation, and this poor-oxygen blood is received by the right atrium through the superior and inferior vena cava. Once received, the right atrium contracts, pumping blood into the right ventricle through the tricuspid valve. The right ventricle pumps the blood to the lungs for oxygenation through the pulmonary artery.

The left side of the heart receives the oxygenated blood from the lungs through the pulmonary veins in the left atria, which pumps it to the left ventricle for further distribution to the entire body through the aorta.

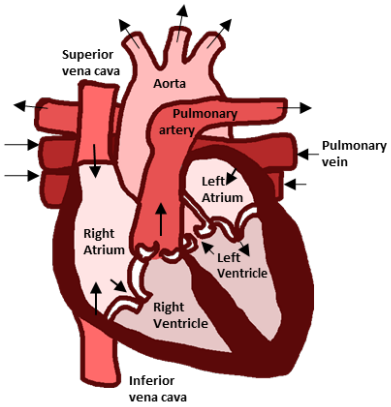

Figure 1.1: Schematic image of the anatomy of the heart.

This cycle of relaxation and contraction of the chambers needs to be constant and synchronous, so the transfer of the blood from one chamber to the other is done sequentially at the maximum strength. This sequence is initiated by spontaneous electrical activity of cells in pacemaker tissue. This impulse is modulated by the autonomic nervous system. The electrical potential difference created by the depolarization and repolarization wavefronts can be measured with electrodes placed on the body surface, resulting in the electrocardiogram (ECG).

1.2.2 The cardiac electric cycle and the ECG

The ECG is a graphical representation of the electrical activity of the heart, where each heartbeat is represented by a sequence of characteristic waves. Each wave corresponds to one of the stages of the electric cardiac cycle, which is depicted in Figure 1.2.

The normal sequence starts with the spontaneous depolarization of the cells in the sino-atrial (SA) node, located high in the right atria. This natural pacemaker creates regular electrical impulses that determine the heart rate and propagate along the conduction system of the heart.

The action potential generated by the SA node travels through the atria causing its depolarization, which is translated into a small wave in the ECG referred to as P-wave. The electrical current is then conducted from the SA to the atrioventricular (AV) node. The electrical impulse is delayed in the AV node, so the atria finishes depolarizing before the stimulation of the ventricles starts. This is visible in the ECG as the iso-electric PQ segment.

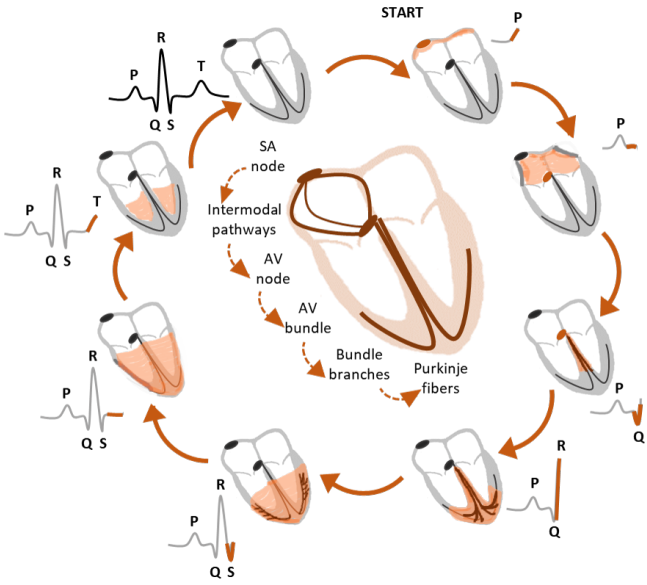

Figure 1.2: Steps of the cardiac cycle and relation with the ECG waves. Figure adapted from [158, 177].

From the AV node, the electrical impulse travels through the bundle of His, the bundle branches and the Purkinje fibers causing fast depolarization of the ventricles, which is visible in the QRS complex. At the same time, the atria is repolarizing, causing it to contract, but this process triggers smaller voltages which are masked by ventricular depolarization. By convention, the different components of the QRS complex are the Q-wave,, which corresponds to the first negative deflection at the start of the depolarization; followed by the maximum positive peak called the R-peak, and finishing with the S-wave, which is the second negative wave at the end of the depolarization. The QRS-complex is followed by a new short segment of electrical inactivity, which is captured in the iso-electric ST segment. Then, the ventricles repolarize and contract, causing the ejection of blood. This process is captured by the T-wave. At the end of this wave, the aortic valve closes, causing the ventricles to relax and get filled again.

The morphology described above corresponds to a normal heartbeat. A sequence of this type of normal heartbeats creates a normal, stable heart rhythm called normal sinus rhythm. Variations from these standard waves represent disruptions in conductivity that can be associated to structural abnormalities of the heart or alterations in the electrical conduction pathways. The morphology presented in Figure 1.2 corresponds to lead II of the ECG, which provides a frontal view of the cardiac electrical activity. As

explained before, multi-lead approaches capture this same activity from different points of view, resulting in different morphologies where still the same ECG waves can be identified.

12-lead ECG

A clinical examination of cardiology often implies the recording of 10-seconds 12-lead ECG. This is measured using 9 electrodes distributed on the thorax and limbs as shown in Figure 1.3. In some ambulatory or noisy set-ups, an additional lead in the right leg is added for common mode rejection.

The 3 limb electrodes are placed on the chest close to the limbs: left and right arm and left leg (L, R, F respectively). These electrodes are used to construct the first six leads of the 12-lead set-up, including the limb leads I, II and III and the augmented leads aVR, aVL, and aVF, as shown in Figure 1.3. These leads provide a frontal view of the electrical cardiac activity.

The other 6 electrodes are the chest precordial electrodes V1, V2, V3, V4, V5 and V6 that result in homonymous leads. These leads are placed around the chest, providing, in combination with the "central terminal of Wilson" (i.e., the average of the RA, LA, and LF voltages, shown in Figure 1.3 bottom right) a transversal view of the cardiac electrical activity. An example of a 12-lead ECG signal can be seen in Figure 1.4. The morphologies presented in each of the leads differ from each other, but they all correspond to normal heartbeats captured from different points of view.

1.2.3 Aberrant patterns in the ECG: abnormal heartbeats and arrhythmia

The identification of aberrant patterns is a key step in ECG analysis, since they can be associated to different pathologies. This analysis can be performed at a rhythm or heartbeat level.

When abnormalities appear at the heartbeat level, it can be due to the formation of an abnormal focal impulse [31]. These can be referred to as irregular heartbeats, and their detection and classification according to their etiology (atrial, junctional or ventricular) is part of the pre-processing stage of ECG analysis. The morphology of the heartbeat is related to the origin of its excitation. For instance, alterations or absence of P-wave are associated to atrial extrasystoles, while changes in the QRS complex or the T-wave are linked to the ventricles.

Sporadic irregular heartbeats do not necessarily imply any cardiovascular risk [93], but they do not represent the normal activation pattern of the heart. Hence, these heartbeats

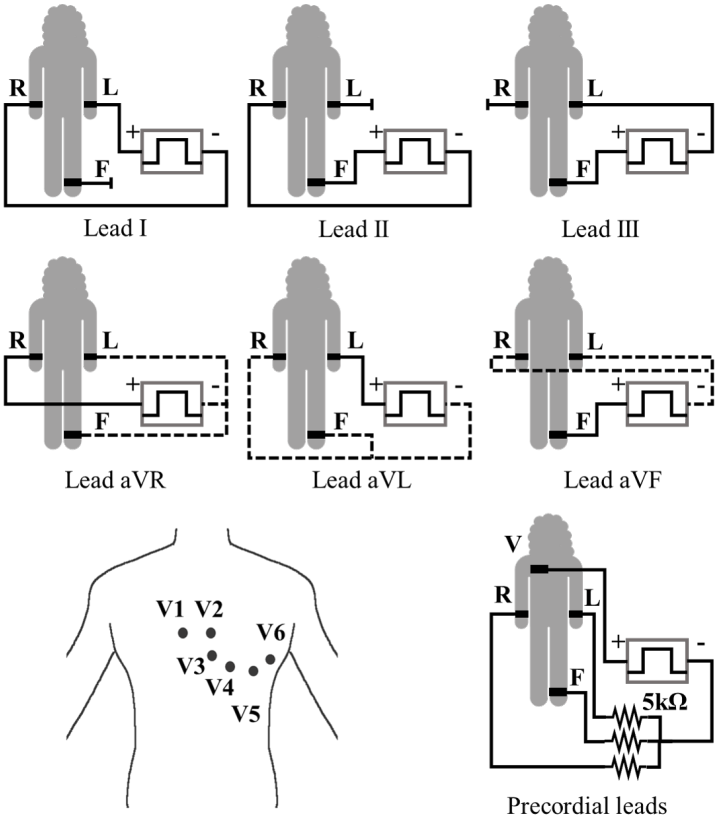

Figure 1.3: Electrode connections to record the standard limb leads (upper row), the augmented limb leads (middle row), and the precordial lead (last row, right). In the bottom left, electrode placement of the precordial leads. Figure adapted from [22].

are normally identified and removed from ECG analysis. This is the case not only for the evaluation of the cardiac activation pattern, but also for heart rate variability analysis (HRV), which must be derived exclusively from intervals between normal heartbeats [110]. In other cases, irregular heartbeats may be included in further analysis but yet identified, since their recurrence and the rate at which they appear can be associated to risks of arrhythmia and cardiac arrest, both in patients with and without structural heart disease [12, 83, 128].

Persistent conduction problems due to abnormal impulse formation or disruptions in normal conduction pathways are called arrhythmia. These can be related to a too slow rhythm, called bradycardia (< 60 bpm), or to a too fast heart rate (> 100 bpm) which

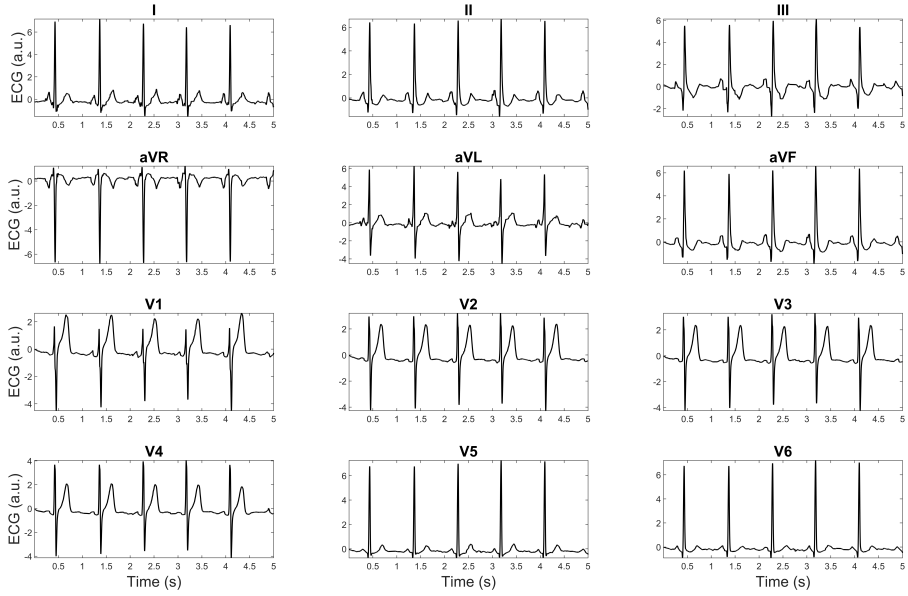

Figure 1.4: Example of a 12-lead ECG signal.

is called tachycardia. Both types of arrhythmia are related with health risks, since bradycardia can cause a lack of oxygenated blood in regions of the heart, which can lead to tissue death; and tachycardia may prevent the heart chambers from getting filled, pumping insufficient blood to the body. Arrhythmia can be atrial, junctional or ventricular, according to their region of origin.

Ventricular tachycardia (VT) is one of the most dangerous rhythm disturbances. It usually is caused by structural heart disease causing fibrosis or scarring of the cardiac muscle, which decreases the contractile capability of the ventricles, mostly the left one. These scars alter the normal activation pathways and can cause reentry of the electrical impulses around them, causing the perpetuation of the fast rhythm. Although less frequently, VT can also appear in subjects without structural heart disease. In these cases, VT is caused by enhanced sympathetic activation of the autonomic nervous systems, which creates spontaneous focal impulses of some part(s) of the ventricle. These fire faster than the SA node, dominating the heart rhythm [137].

VT can trigger ventricular fibrillation (VF), a disorganized fast rhythm in the ventricles that often leads to cardiac arrest and without treatment causes death [119]. VF can appear due to the combination of excessive sympathetic activation together with reduced protective vagal reflexes. The latter can lead to the maintenance of arrhythmia, due to the lack of parasympathetic activity to control these abnormal activations. VT and

VF are the most common causes of sudden cardiac death (SCD), which is the origin of nearly half of the deaths due to cardiovascular diseases worldwide [193]. SCD is defined as death occurring due to cardiovascular causes after 1 hour of the start of the symptoms in a person without previous signs of fatal events [52].

Due to the potentially fatal consequences of VT/VF, the holy grail of cardiology is the prediction of this event in order to ensure treatment for patients at risk. Treatment can include antiarrhythmic drug therapy, ablation or the implantation of a cardiac defibrillator (ICD) [137]. An ICD is a device implanted in the chest that monitors cardiac electrical activity and intervenes when necessary to re-establish a regular heart rhythm. It can pace or induce re-starting high voltage shocks if the natural electrical impulses are insufficient or if the heart falls in a dangerous fast rhythm.

Both for patients with and without an ICD, ventricular arrhythmia prediction is needed since the occurrence of arrhythmia or interventions of the ICD are associated with a higher risk of mortality and hospitalizations and a lower quality of life [47, 137]. Additionally, accurate ventricular arrhythmia, or SCD risk prediction could improve the identification of patients in need of an ICD, which is an open research question. Currently, only 20-30% of the patients with an ICD have required its activation, while the high incidence of SCD indicates that several patients are still missed by the risk prediction strategies [174]. The current implantation criteria is often based on LVEF, which is measured using echocardiography, or the NYHA classification, derived from exercise tests. However, other markers are required to complement these and improve risk prediction models. Some of these markers can be extracted from 12-lead ECG, usually recorded in clinical examinations [119].

1.3 ECG biomarkers for ventricular arrhythmia prediction: previous work

ECG analysis is key in the diagnosis of arrhythmogenic cardiac pathology. Several ECG biomarkers have been identified in the literature, associating their presence to a higher risk of ventricular arrhythmia [99, 150, 151] or mortality [141, 176].

In terms of heart rhythm, HRV and heart rate turbulence (HRT) features have been explored to characterize the autonomic tone. The goal of these features is to describe the balance of the sympathetic and parasympathetic nervous system, as an indication of their coordinated action in the regulation of physiologic processes [56]. The correct functioning of the autonomous nervous system conditions the rhythm and the contractility of the heart. However, ECG signals must be at least 5 minutes long for HRV analysis, with the signal length affecting the reliability of the HRV metrics [149]. On the other hand, HRT is measured after the appearance of a premature

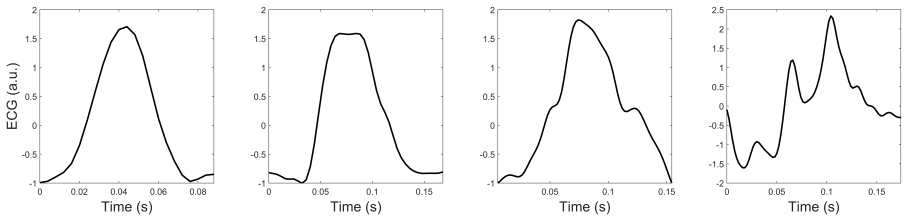

Figure 1.5: Examples of non-fragmented (left) and fragmented QRS complexes. From left to right, fQRS examples with higher level of fragmentation.

ventricular heartbeat or a ventricular paced beat, thus monitoring this in a routine clinical examination is often not possible [14].

Delays in conduction are also relevant arrhythmia risk indicators [6]. Either QRS duration, which captures the time taken for ventricular depolarization, or the time taken between the start of the Q-wave and the end of the T-wave (QT), as the duration of the entire ventricular depolarization and repolarization. Longer times are associated to higher risk of cardiovascular events according to several studies [46, 98, 117].

Another parameter of interest due to its relation with ventricular arrhythmia is repolarization lability, which is the variability of ventricular repolarization in consecutive heart cycles. Higher lability is associated to higher risk of VT, both for patients with and without structural heart disease [11, 70]. This can be quantified measuring the variability of the QT interval, or T-wave alternans [200]. These biomarkers can be cumbersome to extract, since the automatic delineation of the T-wave is a long-standing problem in biomedical signal processing [23, 31].

Other arrhythmia risk-markers can be extracted from the period of ventricular depolarization. QRS fragmentation (fQRS), defined as a QRS with additional complex notches, has been associated to myocardial scar and a higher risk of ventricular arrhythmia and mortality [36, 176]. These morphologies represent conduction disturbances in the ventricles, which can cause sustained arrhythmia as explained in the previous section. Some examples can be seen in Figure 1.5, where a non-fragmented (left) and three fragmented QRS complexes are shown, with increasing level of fQRS. These patterns can be automatically identified in the ECG, but there are different fQRS definitions that can cause some inaccuracies in extraction [109].

None of the ECG-derived biomarkers previously mentioned combines the information captured by the different ECG leads to provide a final metric based on the entire heart geometry. Automatic approaches may identify these risk patterns in each lead independently, but the smart integration of the 12-lead information is done by the cardiologist. While some machine learning methods take these multi-view information

into account [159], most of the models applied to ECG do not focus on the extraction of risk predictors, but on the classification of entire signals. These approaches, despite their potential and outstanding results, may not enable a clear interpretation of the signal analysis process, hampering their applicability in clinical practice [159, 205].

The simplified use of 12-lead ECG information by these algorithms makes them inefficient in terms of data availability, since they do not exploit all the information contained in the multi-lead signals. Additionally, the interpretation of the output of these algorithms becomes cumbersome, due to the high number of variables considered for each of the leads. Therefore, the main motivation of this thesis is to propose advanced high-dimensional techniques for the analysis of multi-lead ECG signals, in particular for the identification of arrhythmia risk predictors. This goal is tackled both for feature-based approaches, by developing an interpretable unsupervised feature selector; and for dimensionality reduction of multi-channel time series, with the development of a novel framework to explore 12-lead ECGs in a lower-dimensional space.

1.4 Research objectives

The aim of this thesis is to propose novel automated approaches for the analysis of multi-lead ECGs, which can holistically consider all dimensions in an interpretable way, and can extract relevant information linked to arrhythmia risk prediction. For that purpose, this work addresses three main research objectives:

1. **Develop robust algorithms for the extraction of arrhythmia risk predictors from the ECG.** The automated extraction of arrhythmia risk predictors presents several challenges, as discussed in Section 1.3. On the one hand, many of these ECG biomarkers are very sensitive to the quality of the ECG [124] and strongly depend on an accurate delineation of the ECG waves, which is a long-standing challenge in ECG processing [23, 31]. This sensitivity is specially challenging in single-lead ECG signals, since their information is limited to a single time series that cannot be enriched by the content of other leads. On the other hand, these algorithms are very dependent on their training and evaluation strategies. This can cause strong variations in performance when applying these algorithms to unseen databases, since their annotations can be affected by inter-observer variability [159, 175]. In this context, the second goal of this thesis is to propose robust algorithms for the extraction of arrhythmia risk predictors. These methods should include strategies to reduce the impact of noise on segmentation and feature extraction; and they must be critically designed to reduce the influence of ambiguous annotations.

2. **Propose a novel framework for dimensionality reduction of 12-lead ECG signals that enables the comparison of different subjects and enhances irregularities.** The biomarkers discussed in Section 1.3 do not combine the information of the 12-leads in the ECG. Since this multi-lead signal provides different views of the cardiac cycle, its redundancy can be exploited while integrating the spatial information contained in each lead for a more detailed analysis of the conduction of the heart. Therefore, the second goal of this thesis is to propose a unified framework to reduce dimensionality of 12-lead ECG signals, enhancing potential abnormalities related to different pathologies and conditions. In order to avoid the potential information loss associated to fixed transformations, often used for instance to derive VCG from 12-lead ECG [188], data-driven methods should be explored. Additionally, this approach should allow the comparison of different subjects in a common framework, in contrast with other data-driven transformations [1, 63].

3. **Develop an unsupervised feature selector based on manifold learning.** Feature-based approaches often lead to high-dimensional data, where every data sample is described by a large feature vector. Feature selectors allow reducing the dimensionality of these data by selecting a subset of the features to capture most of the information contained in the high-dimensional data. When these algorithms are supervised, the selection of features is based on annotations and prior knowledge that are used to evaluate the performance of the selected subset [69]. Nevertheless, biomedical data often lack labels, or if these are available, for some applications these may be subject to inter-observer variability [72, 175]. Therefore, unsupervised feature selectors are often more practical solutions for this task since they can unveil the true underlying structure of the data [69]. Although multiple unsupervised solutions exist, they are often based on complex parametric formulations, and associated with high computation costs [133, 207]. Therefore, the last objective of this thesis is the development of a novel, user-friendly algorithm for unsupervised feature selection, computationally-efficient and which can be made publicly available.

1.5 Chapter overview and personal contributions

This section describes the contribution of each of the chapters to the research objectives of this thesis. A schematic summary of the chapters and their relation with the objectives is shown in Figure 1.6.

Chapter 2. Heartbeat classification in single-lead ECG. A typical ECG analysis pipeline consists of first preprocessing the ECG and then detecting the heartbeats. Once

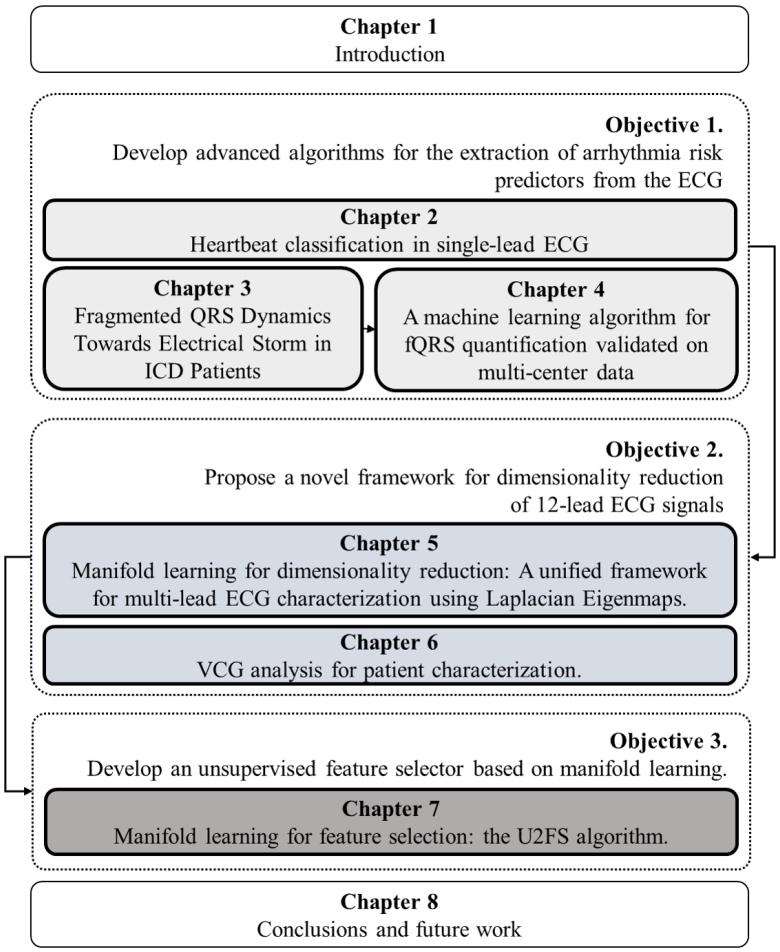

Figure 1.6: Schematic view of this thesis.

this is done, the next step is often classifying these heartbeats according to their origin, as normal, atrial or ventricular. This chapter presents the first contribution of this thesis, namely an algorithm for heartbeat classification in single-lead ECG signals. The focus is to maximally exploit the limited information contained in single-lead signals, while being robust against noise and artifacts for a potential implementation on wearable devices. Additionally, some guidelines are provided for the development of new heartbeat classification algorithms, based on the analysis of the database commonly considered for their training and evaluation: the MIT-BIH arrhythmia database. The first version of this algorithm was presented in the 2018 Computing in Cardiology conference, and it is published in [184]. An improved version of the algorithm and the study focusing on the MIT-BIH arrhythmia database was presented in the annual 2019 symposium of the IEEE Engineering in Medicine and Biology Society, and it is published in [180].

Chapter 3. Fragmented QRS Dynamics Towards Electrical Storm in ICD Patients. This chapter presents the second contribution of this thesis: the analysis of fQRS as an arrhythmia risk predictor in patients with an ICD presenting electrical storm (ES). ES is defined by the appearance of 3 or more ICD shocks in a period of 24 hours. The level of fragmentation in 12-lead signals was quantified using a state-of-the-art algorithm [64]. This algorithm was applied on ECG signals recorded from a group of patients with history of ES and a group of matching controls, over a series of 4 visits. This analysis raised some research questions that motivated the studies included in the next chapters. The results of this analysis were presented in the 2020 Computing in Cardiology Conference, and they are published in [182].

Chapter 4. A machine learning algorithm for fQRS quantification validated on multi-center data. The analysis presented in the previous chapter highlighted several challenges in fQRS quantification. First, the accuracy of such quantification strongly depends on QRS delineation. Second, there is a high influence of inter-observer variability in the identification of fQRS [109, 175]. Therefore, the third contribution of this thesis involves the development of two novel algorithms: one for multi-lead QRS segmentation, which was made publicly available[1]; and one for the quantification of fQRS, validated on multi-center data. This chapter describes these two algorithms and includes a study evaluating the influence of different fQRS definitions in the performance of the fQRS quantification algorithm. This work was submitted to *Scientific Reports* and is currently under review [185].

[1]The multi-lead QRS segmentation algorithm can be found in: https://github.com/avillago/multiLeadSegmentation.

Chapter 5. Manifold learning for 12-lead ECG dimensionality reduction: A unified framework based on Laplacian Eigenmaps. As explained in Section 1.3, most of the algorithms for the extraction of ECG biomarkers associated to risk of arrhythmia do not make use of the 12-leads of an ECG. Either they provide independent outputs for each of the leads, leaving the task of merging the multi-dimensional information to the clinician [64]; or they exclusively consider one of the leads and ignore the rest [125]. Motivated by this gap in ECG analysis, this chapter presents one of the central contributions of this thesis: a unified framework for the analysis of 12-lead ECG signals. This framework is makes use of Laplacian Eigenmaps, which is a non-linear dimensionality reduction method based on manifold learning. The method was applied on the characterization of signals of different groups of subjects, namely healthy subjects; patients with ischemic and non-ischemic cardiomyopathy; and for cardiomyopathy patients with and without history of ES. This work was submitted to the *IEEE Journal of Biomedical and Health Informatics* and is currently under review [181].

Chapter 6. VCG analysis for patient characterization. The previous chapter presents a novel approach for dimensionality reduction of 12-lead ECG signals. This method can be considered data-driven, since the methodology exploits the information contained on the data to reduce dimensionality, without any parametrization. However, already since the 1950s the electrical activity of the heart was graphically described in 3 dimensions using the vectorcardiogram (VCG) [25, 188]. Since VCG can be derived from 12-lead ECG, technically it can be considered a reduced-dimensionality version of the ECG. In this context, and for the sake of completeness, the methodology proposed in Chapter 6 is compared here against VCG formulations. In addition, novel VCG features are proposed and used for the analysis of multi-lead ECG signals recorded from the same populations studied in Chapter 5.

Chapter 7. Manifold learning for unsupervised feature selection: the U2FS algorithm. This chapter presents a review of manifold learning approaches applied to feature selection, and proposes the last personal contribution of this thesis: an unsupervised feature selector based on the utility metric (U2FS). This method is faster and simpler to use than other state-of-the-art algorithms, while obtaining comparable results for different types of data. The U2FS algorithm was made publicly available[2], and it was published in [183].

[2]The U2FS algorithm can be found in: https://github.com/avillago/u2fs.

1.6 Collaborations

This PhD was done under the supervision of professors Sabine Van Huffel, Carolina Varon and Rik Willems. The research activities were conducted in the biomedical data processing research group (BIOMED), which is part of the STADIUS center for dynamical systems, signal processing and data analytics in the Department of Electrical Engineering (ESAT) of KU Leuven. During these years, I collaborated with several people both within and outside the BIOMED group, leading to different conference and journal publications.

The first collaboration within BIOMED was with Sibasankar Padhy, with whom we developed the first version of the heartbeat classification algorithm which was presented in the 2018 Computing in Cardiology conference [184]. This algorithm was improved using feature selection thanks to the collaboration with Margot Deviaene. This work was presented in the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [180]. Both studies are included in Chapter 2 of this thesis.

In my first months of the PhD, I also collaborated with Jonathan Moeyersons, John Morales and Carolina Varon in the annotation of a database for ECG quality. Together we annotated more than 10000 ECGs that were used as ground truth in two journal and one conference publications [28, 123, 124].

Thanks to the essential collaboration with Abhijith Mundanad Narayanan and Alexander Bertrand, we decided to include the utility metric in the development of our unsupervised feature selector U2FS, making it efficient and user-friendly. This work was published in [183] and is included in Chapter 7 of this thesis.

Additionally, this thesis was possible thanks to the strong collaboration with the department of Cardiology of UZ Leuven, which provided most of the ECG signals used in this thesis. The collaboration with Prof. dr. Rik Willems, co-supervisor of this thesis, with dr. Bert Vandenberk and Sebastian Ingelaere, helped me to understand the clinical challenges tackled in this thesis, as well as to interpret the results of our studies.

The first collaboration with Sebastian Ingelaere led to the study on fQRS and ES. This study is included in Chapter 3 and was presented in the 2020 Computing in Cardiology conference [182]. This study triggered the development of our fQRS quantification algorithm and the QRS segmentation method for multi-lead ECG signals. These methods were developed in close collaboration with dr. Bert Vandenberk and Sebastian Ingelaere. The fQRS method was trained and evaluated on the database provided by the multi-center EUropean Comparative Effectiveness Research to Assess the Use of Primary ProphylacTic Implantable Cardioverter Defibrillators (EU-CERT-ICD) project. The analysis of these results was enriched by the insights of Tuomas Kenttä and Heikki V. Huikuri, from the Medical Research Center of Oulu (Finland); Markus Zabel, from

Heart Center, University of Göttingen Medical Center (Germany); Tim Friede from the Department of Medical Statistics of the University Medical Center Göttingen (Germany); Christian Sticherling, from the Division of Cardiology of the University of Basel Hospital (Switzerland); Anton Tuinenburg, from the Department of Cardiology of the University Medical Center Utrecht (The Netherlands), and Marek Malik, from the National Heart and Lung Institute, Imperial College (UK). This work, described in Chapter 4, was submitted to *Scientific Reports* and is currently under review [185].

During the second half of my PhD, the supervision of master thesis students became an important part of my work. I supervised five students during this time. Together with Jonathan Moeyersons, we supervised the thesis of Ben Jacobs, who explored the use of LE to identify deterioration in Tetralogy of Fallot patients. His work was presented in the 2020 Computing in Cardiology Conference (CinC) [85]. His thesis inspired the study presented in Chapter 5 of this book, where we created a unified framework for ECG dimensionality reduction based on LE and we applied it to the analysis of structural and arrhythmogenic abnormalities in 12-lead ECG. This work was done in collaboration with Sebastian Ingelaere, Ben Jacobs and Bert Vandenberk and is currently under review for the *IEEE Journal of Biomedical and Health Informatics* [181]. Together with John Morales, we supervised the thesis of Tinne Briers. Tinne explored the use of VCG derived from 12-lead ECG for the analysis of ES patients. Her thesis inspired the work presented in Chapter 6 of this book. Additionally, I supervised the thesis of Louise Lein, Jef Masereel and Robbert Jacobs.

I would like to acknowledge and thank all the people that I worked with during these years. Their insights and suggestions enhanced the work presented in the following chapters, and they broadened my interests and knowledge on these research topics.

Chapter 2

Heartbeat classification in single-lead ECG

Most of the automated ECG analysis requires a first step of irregular heartbeat detection, since ECG analysis is mostly performed on normal heartbeats during normal sinus rhythm. Additionally, these irregular heartbeats can be classified according to their origin, as ventricular or atrial, exposing and describing abnormalities in the electrical activity of the heart.

This chapter presents an algorithm for automated heartbeat classification in single-lead ECG signals. This algorithm, conceived for ECGs recorded using wearable devices, exploits the temporal and frequency content of the signals to extensively characterize each heartbeat and reduce the impact of noise and other artefacts. Given the long-standing challenge of accurately classifying atrial heartbeats, this chapter also includes an analysis of the dataset commonly used for training and evaluating these algorithms, and the impact of this database on the results. This chapter includes the results published in:

> **Villa, A.**, Padhy, S., Willems, R., Van Huffel, S., & Varon, C. (2018, September). Variational mode decomposition features for heartbeat classification. *2018 Computing in Cardiology Conference (CinC)*, 2018, pp. 1-4

> **Villa, A.**, Deviaene, M., Willems, R., Van Huffel, S., & Varon, C. (2019, July). Are we training our heartbeat classification algorithms properly?. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 6363-6366

2.1 Introduction

Heartbeat classification is a critical step for ECG analysis, since it allows the identification the origin of the heartbeats present in the signal. This classification can be used to detect normal heartbeats, since they describe the common conduction pathways of the heart and therefore, they are the ones included in ECG analysis. Some examples of this are the analysis of heart rate variability [27] or the characterization of ECG morphology to detect fQRS [74]. Additionally, heartbeat classification provides information about the spatial origin of the heartbeats, which can assist the diagnosis of heart diseases if ventricular or atrial heartbeats occur frequently [12, 20].

The classification of heartbeats is highly time consuming and tedious if done manually, specially for ambulatory monitoring. Therefore, there is a big interest in automatic algorithms. Generally, these algorithms include a first step of feature extraction to characterize the heartbeats, and these features are fed into a classifier that discriminates between the different classes. The main difference between normal and irregular heartbeats is their temporal condition, since irregular heartbeats are characterized by arrhythmia, altering the normal sinus rhythm [31]. Furthermore, the morphology must be taken into account to distinguish between the spatial origin of the different irregular heartbeats, either in the atria or in the ventricle. Consequently, the most commonly used features are those related to the temporal differences between consecutive heartbeats [102], morphology [41, 197], wavelets [160], higher order statistics (HOS) [115], or combinations of these [125, 208].

All these approaches can be used on single or multi-lead ECG signals, since they treat each lead independently. However, multi-lead approaches often imply post-processing steps that combine the information from the different leads. This makes these approaches more robust to noise and other artefacts in the signals than for the case of single-lead, since irregularities associated to the cardiac cycle should be visible in more than one lead, while artefacts are lead or location dependent. On the other hand, single-lead analysis poses the challenge of extracting information from a single source, which can be affected by noise masking the relevant waveforms. Additionally, single-lead signals are often recorded with wearable devices, which are more sensitive to noise and artefacts since they record the ECG in an ambulatory way [124]. Therefore, single lead ECG requires specialized algorithms which take these characteristics into account.

The MIT-BIH Arrhythmia database is the most used database to evaluate these algorithms [126]. Typically, this evaluation is performed following the protocol proposed in [41], in order to be able to compare results of different algorithms. Even though several combinations of features and classifiers have been proposed, they often achieve very low performance in the detection of Supraventricular heartbeats (SVEB), due to their difficult distinction from Normal (N) beats in terms of morphology, and the

inter-patient variability within the SVEB class. One of the best methods uses features describing the temporal, morphological and HOS characteristics of the heartbeats, combined with an ensemble of SVM classifiers [125], achieving sensitivities and specificities above 93% for the N and Ventricular (VEB) class, and yet a sensitivity and specificity for the SVEB class of 78% and 50%, respectively.

This study aims to find features that characterize the different classes, improving the current state-of-the-art results for SVEB and making use of single-lead ECG so they can be used on wearable devices. To this end, the use of variational mode decomposition (VMD) [48] for heartbeat classification is explored. This method allows analyzing single-lead ECGs in different frequency bands, expanding the information shown in this single view for a more detailed exploration. While previous works have explored similar methods like empirical mode decomposition (EMD) or wavelet analysis [148, 198], VMD has been proven to be more robust again noise due to its adaptive way of deriving the modes [104]. Therefore, a combination of VMD and temporal features is proposed. Additionally, new suggestions to better train and evaluate heartbeat classification algorithms are suggested to reduce improve performance on the SVEB class.

2.2 Methods

2.2.1 Data

In this study, the first lead of the MIT-BIH Arrhythmia dataset recordings was used for analysis [126]. This public dataset consists of 48 recordings of approximately 30 minutes, sampled at 360 Hz and for which the R-peak locations and the labels of the heartbeats are available. The Association for Advancement of Medical Instrumentation (AAMI) recommends classifying heartbeats in 5 groups: Normal, Supraventricular ectopic beat, Ventricular ectopic beat, Fusion beat and Unknown beat. In this study, these recommendations were followed to make the results comparable to the state-of-the-art [10]. Accordingly, 4 ECG signals were removed from the analysis since they correspond to patients with a pacemaker. In addition, the heartbeat annotations were divided into 5 classes, which are described in Table 2.1.

Given the class imbalance and the heterogeneity of the recordings, the protocol presented by de Chazal et al [41] was used. This protocol suggests the sorting of the recordings of the database in a training and a test set, DS1 and DS2 respectively, in such a way that patients are not mixed between both groups. Furthermore, both groups were designed to present a comparable number of heartbeats of each class. The number of heartbeats of each class, and in both groups, are presented in Table 2.2.

Table 2.1: Types of beats from the MIT-BIH Arrhythmia dataset included in classes Normal,SVEB,VEB,F and Q

	Type of heartbeat	Symbol
N	Normal beat	N
	Left Bundle Branch Block (LBBB)	L
	Right Bundle Branch Block (RBBB)	R
SVEB	Atrial premature beat (APB)	A
	Aberrated APB	a
	Nodal/Junctional prem. beat	J
	Prem./Ectopic supraventr. beat	S
	Atrial escape beat	e
	Junctional escape beat	j
VEB	Ventricular premature beat	V
	Ventricular escape beat	E
F	Fusion ventricular and normal beat	F
Q	Paced beat	P o /
	Fusion of paced and normal beat	f
	Unclassifiable beat	U

Table 2.2: Description of the MIT-BIH arrhythmia database.

	N	SVEB	VEB	Fusion	Unknown
DS1	45850	944	3788	414	8
DS2	44240	1837	3220	388	7

2.2.2 Pre-processing and heartbeat segmentation

All ECG signals were pre-processed to remove the baseline wander and high frequency noise affecting further steps of feature extraction. The method proposed in [41], based on a two-step application of median filters, was used to remove the baseline wander. Firstly, a median filter of 200ms was used to extract the QRS complexes and the P waves, followed by a 600ms median filter to remove the T waves. The remaining signal was then subtracted from the original in order to remove the baseline wander. To remove high frequency noise, a Butterworth low-pass filter with cut-off frequency at 50Hz and order 6 was applied.

The R-peak annotations provided by the MIT-BIH database were used for segmentation, applying a fixed symmetric window of 500ms around the R-peak.

2.2.3 Feature extraction

In this approach, each heartbeat is described by its temporal characteristics, related to the distance *RR* to the R-peaks around each heartbeat, and its frequency and morphology by applying VMD. The first group of features are expected to distinguish between normal and irregular heartbeats, while VMD enhances the differences between each of the irregular classes. The complete list of features is provided in Table 2.3.

Table 2.3: List of features for heartbeat classification.

Time features	Pre RR
	Post RR
	|Pre RR - post RR|
	mean(previous 10 RR) - preRR
	angle post RR, pre RR
VMD features	Bandwidth of the mode
	Power in the band
	Symmetry around the R-peak
	Number of zero crossings
	Difference between minima around R-peak
	Central Frequency
	Skewness
	Variance

Temporal Features

Given the relevance of rhythm in irregular heartbeats, five standard temporal features are included in this work: the time to the previous R-peak, to the next one, the absolute difference between the previous and posterior RR intervals, the deviation of the previous RR interval to the preceding 10, and the angle between post and pre RR intervals. These temporal features are listed in Table 2.3, and they were normalized per patient, as recommended in [102].

VMD features

Variational Mode Decomposition (VMD) is a method to adaptively decompose a signal into its principal modes. Those modes are selected considering that they must be able to reconstruct the signal in an optimal way and band-limited around an automatically estimated center frequency [48].

This decomposition makes the method robust against noise, since the principal modes of the signal characterize the different ECG waves due to the quasi-periodic characteristics of the ECG, and other sources would not contain enough energy as to be captured by the VMD modes. Additionally, the decomposition in a number of k modes enriches the information of single-lead ECG, enabling a multi-scale analysis. Being a parameterized method, VMD requires two parameters to be fixed in advance: the number of modes k and the bandwidth control parameter α. For ECG, the optimization technique developed by Guo et al. [68] leads to a number of modes k of 5 and α of 30. An example of the three most characteristic classes N, SVEB, and VEB and their decomposition arranged in increasing order of central frequency are shown in Figure 2.1. As shown in the figure, the number of zero crossings and the amplitude difference between the deepest minima seem to differ among the three classes, while the symmetry around the R-peak is specially considered to differ between Supraventricular and Ventricular heartbeats.

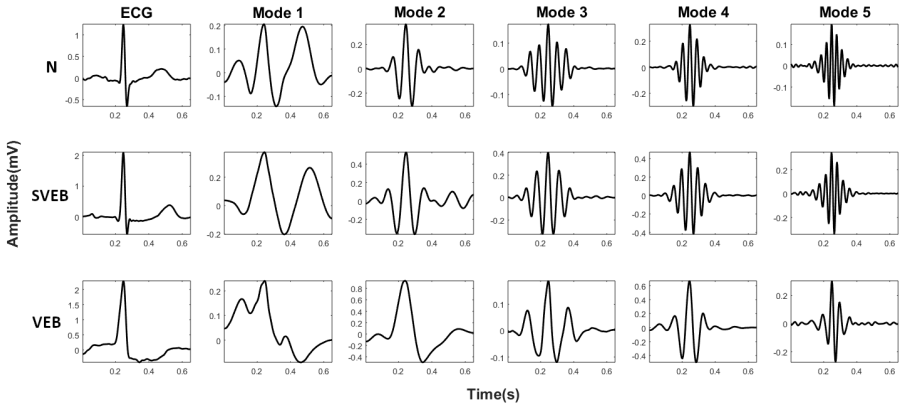

Figure 2.1: Example of Normal (N), Supraventricular (S) and Ventricular (V) beat: ECG and the 5 modes obtained.

Each of the modes is characterized by 8 features, namely the bandwidth of the mode, the power, the symmetry around the R-peak, the number of zero crossings, and the difference between minima around R-peak. In a first version of the algorithm the modes obtained from each heartbeat were arranged in terms of their central frequency, extracting features only from those modes related to the characteristic frequencies of the ECG (from 1 to approx 25 Hz). However, in the final set of features, this restriction was eliminated in order to exploit the frequency differences between each of the classes. Furthermore, by not sorting the modes, the relevance of each frequency band in each of the heartbeat classes is preserved. This new condition is described by the central frequency of each mode as a new feature. Additionally, the skewness and the variance of each mode were also used as features, motivated by the relevance of HOS in heartbeat classification [115]. Therefore, each mode was characterized by the

8 features listed in Table 2.3, making a total of 40 VMD features per heartbeat, plus the 5 temporal features previously described.

2.2.4 Feature selection

Given the high number of features used to describe each heartbeat, a stage of feature selection was introduced in this study. A backwards wrapper feature selector was used, based on random forest as classifier and the change in the Cohen's Kappa κ value as performance evaluation metric, given the unbalance in the class distribution [45]. This supervised method performs classification using the entire feature set and evaluates the difference in performance using the κ index when the values of the feature of interest are randomly permuted. Those features affecting less the performance are assumed to be redundant or irrelevant, and they are removed from the feature set. This process is repeated until only one feature remains for classification.

2.2.5 Classifier and performance metrics

The selected features were fed into an LS-SVM classifier using an RBF kernel, using 10-fold cross-validation for tuning the parameters of the kernel [164].

The classification results were derived from the confusion matrix as suggested by the AAMI standards, with the peculiarity of not rewarding or penalizing the classification of Fusion beats as VEBs. In addition to the sensitivity and specificity of each class, general performance metrics that take into account the high imbalance of the classes must be used. 3 indexes were considered for this aim: (1) the j index, which adds the sensitivities and specificities of the most important arrhythmias (SVEB and VEB) [160]; (2) the Cohen's Kappa κ, as a measure of agreement between the ground truth and the classifier; and (3) the jk index [115], which is a weighted sum of the j and κ indices.

2.3 Results and discussion

2.3.1 Selected features

The Cohen's Kappa κ value extracted from the feature selector is depicted in Figure 2.2. This error provides information about the relevance of the features. Since the drop of the classification error starts around 35 features removed, the last 10 features were

kept for further classification. The features selected are listed in Table 2.4 in decreasing order of relevance. The features related to VMD are presented in bold.

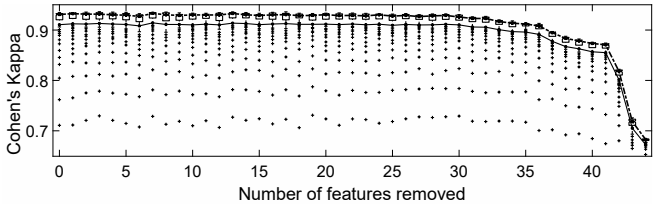

Figure 2.2: Classification error vs. number of features removed

Table 2.4: Selected features in order of relevance

Pre RR	Time
Mode 1 Variance	VMD
Mode 1 Central Freq	VMD
angle postRR,preRR	Time
PostRR	Time
Mode 1 Power	VMD
mean(previous 10 RR)-preRR	Time
Mode 2 Central Freq	VMD
Mode 2 Skewness	VMD
Mode 1 Symmetry mode around the Rpeak	VMD

As it can be seen, 4 out of the 5 proposed temporal features were selected, given the relevance of rhythm in the detection of irregular heartbeats. The angle between the post and pre RR intervals, is the second most important temporal feature. Furthermore, the comparison of the previous RR interval with the tendency of the previous 10 intervals shows more relevance than the difference between previous and posterior RR intervals. Regarding the VMD features, it is clear that the most relevant modes are the first ones, generally related to the lowest frequencies. The new features, related to HOS and to the central frequency of the first and second modes, are among those selected, representing the distinction in frequency content of the different classes.

2.3.2 Classification results

Table 2.5 shows the classification results of our method, with the step of feature selection that reduces them to 10, and it compares them with the state-of-the-art. The results are presented for each of the classes independently, and jointly in different unifying metrics. The sensitivity of our method for the SVEB class is at the level of

Table 2.5: Results on MIT-BIH (DS2) comparing this study to state-of-the-art methods.

	Perf. Metrics	Our method	Chazal et al[41]	Classifier Mar et al[115]	Zhang et al[208]	Mondéjar et al[125]
N	**Se**	*0.82*	0.87	0.9	0.89	0.96
	Sp	*0.98*	0.99	0.99	0.99	0.98
SVEB	**Se**	*0.83*	0.76	0.83	0.79	0.78
	Sp	*0.29*	0.39	0.34	0.36	0.5
VEB	**Se**	*0.88*	0.8	0.87	0.86	0.95
	Sp	*0.84*	0.87	0.76	0.93	0.94
F	**Se**	*0.11*	0.89	0.61	0.94	0.12
	Sp	*0.08*	0.09	0.17	0.14	0.24
All classes	Se	*0.63*	0.83	0.8	0.87	0.7
	Sp	*0.53*	0.57	0.56	0.6	0.66
	Acc	*0.81*	0.86	0.9	0.88	0.95
	***j*-index**	*2.83*	2.77	2.8	2.93	3.17
	κ score	*0.5*	0.53	0.6	0.59	0.76
	***jk*-index**	*0.61*	0.61	0.65	0.66	0.77

Se: Sensitivity; Sp: Specificity; Acc: Accuracy

Mar et al. [115], being the best among the state-of-the-art. Regarding the sensitivity of VEB, this study presents the second best results, only below [125]. However, in terms of specificity, neither SVEB nor VEB class reach the best results of the literature, affected by the miss-classification of some of the N beats. The F class, on the other hand, presents very low sensitivity. This can be explained by the design of the features for this analysis, which originally aimed to tackle specifically SVEB and VEB.

2.3.3 Feature analysis

The analysis of the results presented in Table 2.5 highlights the general problem of the detection of SVEB in all classifiers. This problem has commonly been explained by the similar morphology of N and SVEB, even though temporal features should clearly differ from the N class.

Figure 2.3 shows the kernel function estimate of (a) the previous and (b) posterior RR interval for the SVEB class in DS1 and DS2. These temporal features were normalized as suggested in [102], and they were selected because they are often used in literature for heartbeat classification.

As it can be seen, the DS1 and DS2 distributions for the previous RR interval do

not overlap, which explains why a classifier trained with DS1 is not able to properly classify DS2. Furthermore, there is a clear peak in DS2, corresponding to the record 232. 78% of the heartbeats in this record correspond to SVEB, which represent 50% of the SVEB in the entire database, and 77% of the SVEB in the test set.

93% of the SVEBs present in DS1 come only from 5 patients, and 40% of these belong to one single patient. Similarly, regarding the F class, 91.5% of these beats in the entire database belong to 2 patients, one used for train and the other for test in the given protocol.

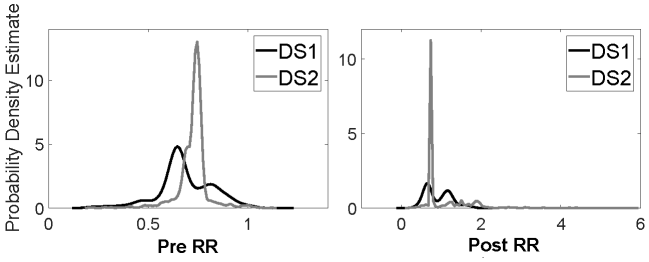

Figure 2.3: Kernel function estimate for RR intervals of the SVEB class

The described differences between patients and the unbalance of the class distribution emphasize the role of the database in the poor results generally obtained in the state-of-the-art for the SVEB and F classes. Given the scarcity of well-known substitute databases, a solution could be a different selection of patients as training set, which could lead to better training of heartbeat classification algorithms.

2.3.4 Alternative results

A new study was performed, using the same 45 features, but dividing the recordings in a new training and test set. Nevertheless, this distribution was done taking into account the amount of recordings present in each group, and the presence of all classes approximately comparable in both datasets. After dividing the files, feature selection was applied to the new training set and those features were used to describe the heartbeats. The results obtained for the N, SVEB and VEB classes are shown in Table 2.6 for 3 cases: (1) DS1 and DS2 as recommended in [41], (2) using DS2 as training and DS1 as test and (3) a new random setting suggested.

These results show that there can be a wide variability in performance, especially for the SVEB class, if the recordings are shuffled differently to the well-known protocol. Even though the amount of SVEB samples and the diversity of their recordings are not enough to properly characterize this phenomena, the distribution of the recordings among train and test set seem to additionally play a major role.

Table 2.6: Classification results for different training-test sets

	N		SVEB		VEB	
	Se	**P+**	**Se**	**P+**	**Se**	**P+**
(1)	0.82	0.98	0.83	0.29	0.88	0.84
(2)	0.67	0.99	0.87	0.35	0.78	0.87
(3)	0.67	1.00	0.83	0.41	0.89	0.66

Given the amount of combinations that could be considered for the redistribution of recordings, future work should explore methods that can automatically select the most representative patients of a database, in order to obtain an optimal training set, where the majority of the examples can be presented to the classifier. This would be useful for this and many other biomedical applications, which require patient independent validation for a robust evaluation.

2.4 Conclusion

This chapter proposed the use of Variational Mode Decomposition (VMD) as a source of features for heartbeat classification. A combination of VMD and time features was proposed to train a classifier based on the MIT-BIH arrhythmia database. The relevance of the features related to the first variational modes and the temporal features was identified by a backwards wrapper feature selector. The selected features were classified using LS-SVM to distinguish 4 heartbeat classes suggested by AAMI: Normal, Supraventricular, Ventricular and Fusion. This setting achieved results aligned with the state-of-the-art, which yet presents poor results for the SVEB class.

Future work should consider the addition of specific features to distinguish the F and SVEB classes. Furthermore, the cascading of classifiers, as suggested in [125], should be studied, aiming to focus on rhythmic heartbeats vs arrhythmia and subsequently on the distinction between arrhythmias.

An analysis of the temporal features and of the patients present in the database was provided, emphasizing the role of the database in the results obtained in the literature. Methods that allow the automatic selection of the most interesting patients for training classifiers are required, or as an alternative, new heartbeat databases that provide more representative samples of the critical heartbeat classes.

Chapter 3

Fragmented QRS Dynamics Towards Electrical Storm in ICD Patients

Electrical storm (ES) in ICD patients, defined as 3 or more appropriate ICD interventions within a time span of 24 hours, is a medical emergency associated with adverse outcome. However, it is debated if ES is only a marker of progressive near end-stage cardiac disease or an arrhythmogenic entity on its own. Better understanding and prediction are necessary to manage the burden of ES.

This chapter explores the relationship between the presence of fragmented QRS (fQRS) and the manifestation of electrical storm in patients with an ICD for ischemic heart disease. The algorithm proposed in [64] was applied to a multi-visit registry of 100 patients with an ICD, where half of them presented ES in the last visit. This preliminary study raised the research questions that triggered the studies in the following chapters. The current results were presented in:

> **Villa, A.**, Ingelaere, S., Van Huffel, S., Willems, R., & Varon, C. (2020, September). Fragmented QRS Dynamics Towards Electrical Storm in ICD Patients. *2020 Computing in Cardiology Conference (CinC)*, 2020, pp. 1-4.

3.1 Introduction

Electrical storm (ES) in ICD patients is defined as 3 or more appropriate ICD interventions within a time span of 24 hours. This medical emergency generally leads to adverse outcome [50, 58]. ES is a difficult to predict dramatic condition with increasing prevalence [78]. Current research still debates if ES is only a marker of progressive near end-stage cardiac disease or an arrhythmogenic entity on its own. The underlying pathophysiological mechanisms are not yet fully understood and require further investigation. Given the fact that ES is defined as a recurrent arrhythmia, current tools for arrhythmia prediction can be taken into account to develop a predictive strategy. One of the most known markers of arrhythmia is the presence and level of fragmentation in the QRS complexes [37]. Furthermore, the development of better predictive models is necessary to reduce its incidence and burden.

Electrical storm (ES) in ICD patients is a devastating arrhythmic event defined by the occurrence of 3 or more appropriate ICD interventions within a time span of 24 hours. This medical emergency has important consequences on the patients quality of life, morbidity and even life expectancy[50, 58]. However, it is debated if ES is only a marker of progressive near end-stage cardiac disease or an arrhythmogenic entity on its own. Better understanding and prediction are necessary to manage the burden of ES. By assessing the risk of developing ES, high-risk patients could receive a specific care pathway with, among others: more regular follow-up, more intense medical therapy and a lower threshold for alarm settings of telecardiologic device monitoring. The aim of this thesis is not to predict imminent arrhythmia, although this might have also possible implications within ICD-therapy, but rather to detect subtle long-term changes in time reflecting the development of a substrate vulnerable for ES. For this aim, ECG-based arrhythmia risk predictors may be analyzed in the development of ES. One of the most known markers of arrhythmia is the presence and level of fragmentation in the QRS complexes [37, 39, 176].

The aim of this work is to explore the relation between fQRS and the manifestation of electrical storm in patients with an ICD. In order to automatically quantify the level of fragmentation in ECG signals, the algorithm proposed by Goovaerts et al. was used [64]. The fQRS scores of each lead are compared between a group of patients presenting ES and a matched control group. Additionally, the evolution of the scores over time is studied for each of these populations.

3.2 Methods

The ES patients were recruited from the database of patients with an ICD in the University Hospitals Leuven (UZL). This is a single-center registry containing

information of all patients with ischemic heart disease (IHD) or dilated cardiomyopathy (DCM) implanted with an ICD between 01/01/1996 and 31/12/2018. The Ethical Committee of UZ Leuven approved retrospective analysis on this database. This database contains 12-lead ECG signals of 10 seconds, sampled at 250, 500 or 1000 Hz. Two subsets of these patients were included in the analysis: 50 patients (94% male, 63±10 years, 37.3±14.3% LVEF at implant) with ischemic cardiomyopathy presenting with ES were identified. Baseline demographics and raw 12-lead ECG data were collected for these patients from the time of implant until storm. As control subjects, 50 patients (94% male, 62±12 years, 34.2±12.2% LVEF) without history of shocks and from the same registry were included.

To enable an analysis of these patients over time, 4 visits were considered for each ES patient: (1) from 36 to 18 months pre-ES, (2) from 18 to 6 months pre-ES, (3) from 6 months before to days before the storm and (4) from 0 to 48 hours after the storm. For the control group, 4 visits similarly spaced in time were considered.

3.2.1 Preprocessing and Segmentation

Baseline wander and high frequency noise were removed from each lead of the ECG using a high-pass and a low-pass fourth-order Butterworth filter, with 0.5 and 70Hz of cut-off frequencies, respectively. Following [64], these filters were applied bidirectionally to obtain zero-phase distortion, and each lead was standardized using z-score to remove amplitude differences between recordings. Recordings presenting ventricular pacing or arrhythmia, such as atrial fibrillation or ventricular tachycardia, were manually removed from the study, since the algorithm used to quantify fragmentation did not consider such segments in its training [64]. Table 3.1 summarizes the amount of records available for each visit after the removal of recordings presenting arrhythmia.

Table 3.1: Number of subjects present in each visit

	Months pre-ES			**Time post-ES**
	36 to 18	18 to 6	6 to 0	0 to 48h
ES Patients	28	34	36	28
Controls	30	36	23	30

The segmentation of the heartbeats was done semi-automatically to exclude the influence of noisy fluctuations around the QRS. The QRS complexes were automatically detected using the R-DECO software [122]. Visual inspection was further performed to remove irregular heartbeats and noisy segments. These corrections were done for lead II. The fiducial points in this lead were automatically translated to the rest of the leads finding maxima in a fixed-size window around the R-peaks in the absolute signal. Once

the R-peaks were detected, the QRS complexes were segmented using a symmetric window of length $\sqrt{RR}$ around the R-peak, as suggested in [31].

The last stage of preprocessing consisted of automatically selecting the usable heartbeats in each 10-second segment for each lead based on cross-correlation. Since only lead II was visually inspected, the QRS complexes in other leads might still present artifacts or noise. Therefore, a stage of noise removal was applied to keep only the cleanest heartbeats for each lead. In this stage, each segmented QRS complex is compared with the rest of the heartbeats present in that record by means of normalized cross-correlation. Only those heartbeats with a correlation above a quality limit were selected. By visual inspection, this limit was set to 0.7. If only less than half of the heartbeats were kept for the analysis, the segment was directly rejected. This allowed to remove extra noise that could lead to errors in the fQRS quantification.

3.2.2 Feature extraction

The level of QRS fragmentation of each segment was automatically quantified using the algorithm in [64]. This model derives 10 features from the ECG signal using Variational Mode Decomposition (VMD) and Phase-Rectified Signal Averaging (PRSA).

VMD and ECG features

Similarly to the study presented in the previous chapter, VMD was used in [64] to extract features, this time to identify signs of fragmentation. Using the parameters defined in [64] to apply VMD to ECG signals, the number of modes k was set to 5. Since fragmentation presents as additional high frequency components in the ECG signal, the features derived from VMD for fQRS quantification were obtained from the modes corresponding to the highest frequencies of the signal, i.e. u_3, u_4 and u_5. The features defined to capture fragmentation were the number of local optima per segmented QRS and the central frequencies ω_k of the 3 highest modes, leading to 6 VMD-based features.

An additional feature was directly derived from the ECG, corresponding to the number of local peaks in each QRS.

PRSA features

The additional 3 features were obtained from PRSA. This technique is used to detect quasi-periodic oscillations in non-stationary signals. While it had been used to evaluate the adaptation capacity of the heart by applying it to the tachogram [152], the work in [66] applied it directly to the ECG to characterize fragmentation.

PRSA consists of 3 steps. Firstly, a set of anchor points is selected, in the increasing and the decreasing part of the QRS complex. A modification was here applied to the original method for feature extraction, to ensure the reduction of the influence of oscilations potentially adjacent to the QRS, often related to noise. In order to reject the increases and decreases potentially present in the signal at the sides of the QRS, a mask was applied to the QRS. This mask has 2 different regions corresponding to the center and the sides of the QRS, dividing the length of the QRS as 20-60-20. A different search method is applied to each of the stages of the mask. In the center of the mask, the selection of anchor points is made using the derivative of the signal, as proposed in [64]. This allows to sensitively detect segments of increase or decrease of the signal, which is required to find the short high-frequency fluctuations corresponding to fragmented QRS. In the sides of the segmented QRS, a less sensitive technique is applied to select the anchor points. This is done by constructing a derivative based on the average of the previous s samples to the anchor point, setting s to 5% of the complete length of the segment.

The second stage consists of applying a window around the anchor point of length $2L$, with $L = 20ms$. Finally, all segments are averaged to obtain the PRSA curve, from which the 3 features are derived. It has been shown in [66] and [64] that the averaged segment of fragmented QRS examples is flatter than that of healthy QRS segments. Thus, the 3 features derived from the average segment are the slope of its linear fit, the magnitude of the crossing of the linear fit with the y axis, and the mean derivative of the obtained segment.

3.2.3 Automatic quantification of QRS fragmentation

Once the 10 features were extracted, they were fed to the model described in [64], which assigns a score between 0 and 1 to each lead of each ECG signal. This model was built using a Support Vector Machine classifier (SVM) [33], which was trained with 10 seconds ECG signals from UZ Leuven. These signals were annotated by 5 experienced cardiologists as 0 or 1, for the absence or presence of fragmentation respectively [175], according to the criteria proposed by Das et al. [36, 38]. The training samples were finally labeled from 0 to 5 corresponding to the addition of the scores given by the cardiologists. This resulted in higher scores for those QRS which were clearly fragmented, and lower scores for non or more doubtfully fragmented

segments. Platt scaling was used to transform the binary output of the classifier into a continuous scale from 0 to 1, using logistic regression [140].

For the data presented in this study, the classifier was applied to each of the visits available for each patient, leading to a set of 12 scores per visit per patient. A set of 4 global scores were derived for each of these visits: the average of the 12 leads, the median, the standard deviation and the maximum score obtained.

3.2.4 Statistical Analysis

Two hypotheses were formulated to see the relation between fQRS scores and the manifestation of ES. For the first hypothesis, a difference in the level of fragmentation between patients presenting ES and control patients was expected. This was evaluated using the Wilcoxon–Mann–Whitney test for each of the 3 global scores derived from the data in the last visit of the patients, corresponding to the moment after storm. In particular, the levels of fragmentation were hypothesized to be higher for the ES patients than for the controls.

The second hypothesis proposed aimed to find trends over time for each of the groups. By comparing the global scores between the first and consecutive visits, different trends were expected to be found for each of the groups. This test was evaluated using the Friedman test for paired data. Therefore, only the patients who had records for both the first and the last visit were included in this test.

3.3 Results

3.3.1 Differences between ES and control patients in the last visit

Figure 3.1 shows the boxplots of both groups in the last visit for the different global scores considered. ES patients are referred to as *ES* and control patients as *controls*. A slight difference can be observed between both groups, generally indicating a higher global fQRS score for ES than controls, together with a higher standard deviation. This is aligned with the first hypothesis. Nevertheless, the Wilcoxon test did not show significant differences between both groups in any of the considered global metrics.

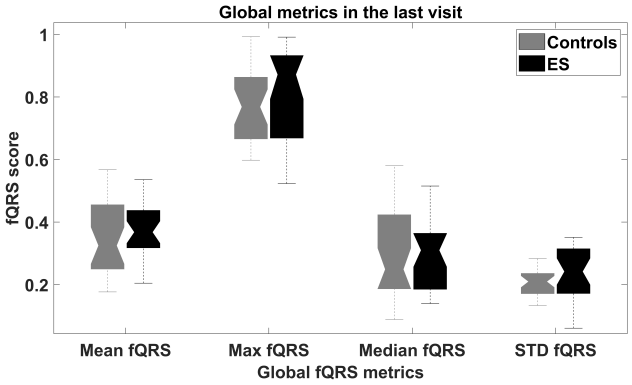

Figure 3.1: Global fragmentation scores for the last visit of ES patients and control patients.

3.3.2 Differences between the evolution of ES and control patients

Figure 3.2 shows the boxplots of both groups from the first to the last visit around ES for the average fQRS score. An increasing trend can be observed in the ES group, which is not visible in the controls. The highest scores are obtained in the third visit, around 6 months before the storm, and these values remain for the last visit.

In order to confirm this observation, a Friedman test was performed. This was done to compare the average score of the first and the last visit, for each of the groups. Since this is a test for paired data, only the patients who had records for both visits were included in this test. This led to 28 and 30 ES and control patients, respectively. While the control patients obtained a p-value of 0.67, the ES patients show significant differences with a p-value$<$0.05.

3.4 Discussion

The metrics proposed to integrate the 12 scores of the last visit did not identify significant differences between ES and control patients. This does not correspond with the first hypothesis, since the ECGs of ES patients in the last visit should correspond to their moment of maximal deterioration, and hence, to the biggest differences between both groups. However, the results of the second experiment suggest that there is a trend towards deterioration in those ICD patients that presented electrical storm, which is not present in the control patients. This is reflected in a higher average fragmentation score

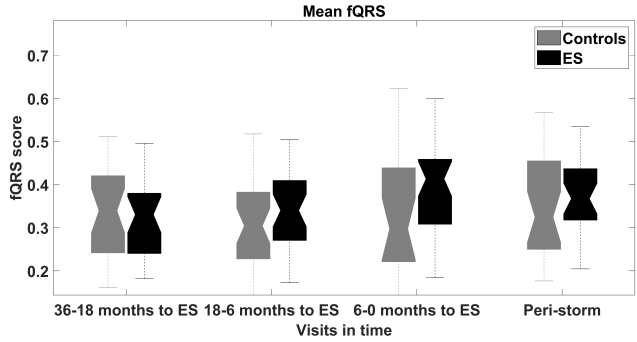

Figure 3.2: Average fragmentation score along the visits, from 3 years before to the time around electrical storm.

of the leads. This is aligned with the literature, that suggests fQRS as an indicator of potential arrhythmia [37], which could eventually become ES. This confirms the need to closely monitor the level of QRS fragmentation in ES patients.

Nevertheless, this study presented some limitations, both in terms of methodology and data. First, the use of an automatic tool for fQRS quantification may lead to inaccurate results. Given that fQRS quantification is a challenging task even for experienced cardiologists, the use of an algorithm trained exclusively on a single dataset with a single fQRS definition [36, 38] can just replicate errors introduced by visual inspection. Additionally, a correct segmentation of the QRS complex is crucial for the proper quantification of fQRS. In a first exploration of the ES data, the complete methodology proposed in [64] was used for fQRS quantification. This methodology included a QRS segmentation algorithm based on VMD, which based on visual inspection of the results obtained, lead to recurrent wrong delineations, specially in cases with noise but also on simpler cases. Those errors motivated the design of a new segmentation strategy, which is the one applied in this work. Nevertheless, the proposed methodology strongly relies on lead II for delineation. While this simplifies the approach, it makes it too susceptible to the quality of this lead, while the rest of the leads can provide additional information to obtain more robust results. These two methodology limitations, namely the QRS segmentation algorithm and the inclusion of additional datasets in the training and evaluation of an algorithm for fQRS quantification will be tackled in Chapter 4.

Additionally, the algorithm used for fQRS quantification provides an independent score for each lead, while clinical analysis would benefit from a global score that allows to characterize patients with a single metric [176]. The definition of this global score requires further investigation, as well as the relation of new regional scores with the type of implanted ICD or regions of ischemia.

Finally, the dataset considered also posed a challenge for the present analysis. The visual analysis and exclusion of recordings presenting arrhythmia or noise was performed without the supervision of clinicians. Additionally, the complete dataset was not visually inspected, i.e. not all the leads were supervised, and therefore noisy signals may have been considered in the analysis, leading to biased or erroneous results. Finally, the exclusion of certain recordings led to missing data for some of the patients, which hindered the analysis.

In Chapters 5 and 6, two approaches for ECG dimensionality reduction will be discussed, tackling the limitations mentioned above.

3.5 Conclusion

A study exploring the relation between QRS fragmentation and the manifestation of ES in patients with ischemic heart disease and ICD is presented. The dynamics of QRS fragmentation in 50 patients with a history of ES were compared to 50 control patients.

The quantification of fQRS was done automatically using a state-of-the-art algorithm, and the results were evaluated over time. A trend towards fQRS deterioration in patients presenting ES was discovered from around 3 years before the event. While this work presents some limitations, the results of this study suggest the relevance of fQRS in the presented group of patients as a potential biomarker of the risk of ventricular storm.

The following chapters tackle the improvement of the QRS segmentation and fQRS quantification algorithms (Chapter 4), and the characterization of ES and non-ES ECGs using methods that combine the information contained in all 12 leads of the ECG (Chapter 5 and 6).

Chapter 4

A machine learning algorithm for electrocardiographic fQRS quantification validated on multi-center data

The previous chapter suggested some aspects that could be improved on automated methods for fQRS quantification. Both visual and automated fQRS analysis is often binary, where heartbeats are annotated simply as fragmented or not. This analysis is prone to intra- and inter-observer variability, given the lack of standardized definitions of fQRS. Therefore, objective quantification of fQRS is needed and could further improve risk stratification of these patients.

This chapter presents an automated method for fQRS detection and quantification. Additionally, a novel robust QRS complex segmentation strategy is presented, which combines multi-lead information and automatically excludes abnormal heartbeats. This methodology was trained to discriminate fragmented from non-fragmented ECG-traces using multi-center data and combining different fQRS criteria used in clinical practice. This work was submitted for publication as:

> **Villa, A.**, Vandenberk, B., Kenttä, T., Ingelaere, S., Huikuri, H. V., Zabel, M., Friede, T., Sticherling, C., Tuinenburg, A., Malik, M., Van Huffel, S., Willems, R. & Varon, C., A machine learning algorithm for electrocardiographic fQRS quantification validated on multi-center data. *Under review in Scientific reports.*

4.1 Introduction

The presence of fragmentation in the QRS complex on a 12-lead ECG has been reported to be an indicator of myocardial scarring or fibrosis [36]. fQRS has been shown to be related to arrhythmic and mortality risk in patients who received an ICD in primary prevention of sudden cardiac death [176]. Therefore, fQRS is a biomarker of interest in risk stratification of cardiac patients. Currently, the presence of fQRS is diagnosed by visual interpretation of the 12-lead ECG to identify irregularities within the QRS complex. This requires specific training to distinguish fragmentation from other patterns, such as early repolarization or non-specific conduction delays. Different definitions of fragmentation, which vary slightly in some of the morphologies considered, have been described. The most frequently used criteria are these proposed by Das et al. [36, 38], which define morphologies which should be identified as fQRS, both for narrow ($\leq$120 ms) and broad ($\geq$120 ms) QRS complexes. However, the application of these criteria might be ambiguous and might lead to intra- and inter-observer variability [175]. Therefore, other criteria have been proposed to overcome these uncertainties, such as those described by Torigoe et al., 2012, where risk stratification was based on the number of fragmented leads [172], or by Maheshwari et al., 2013, where an automatic approach is proposed to detect different patterns of fQRS [108]. However, these methods have neither been validated nor do they express the severity of fQRS. More recently, Haukilahti et al. defined benign and malignant fQRS variants in an attempt to improve the specificity of fQRS as a biomarker [74].

As the application of different fQRS criteria and inter-observer variability limit the reproducibility of clinical studies in different centers [109], the detection of fQRS may benefit from an automatic signal-based, algorithmic approach, to provide decision support. However, there is no true "golden standard" and automatic algorithms trained using predefined fQRS criteria might reproduce the same ambiguities as human annotations. Some automatic fQRS detection algorithms have previously been described most of which are based on signal decomposition methods, such as wavelet transform or intrinsic time-scale decomposition for the delineation of the QRS complex and detection of fragmentation [21, 89, 108]. However, all these algorithms were trained based on their own fQRS criteria in single-center data. The most recent automatic algorithm published by our group (Goovaerts et al., 2018), was based on the Das criteria for fQRS detection [64]. This model, developed in a single-center cohort of sinus rhythm ECGs including narrow and broad QRS complexes, offers a continuous scoring output between 0 and 1 for fQRS quantification and is based on VMD [48] and PRSA [13]. Although this algorithm does not distinguish different fQRS patterns, the output was associated with agreement between observers and the risk of all-cause mortality [65].

This work proposes a fully automated algorithm for fQRS detection and quantification in order to avoid subjective interpretation. In line with this objective, three main

contributions are proposed. First, we present a novel and robust QRS complex segmentation strategy which combines multi-lead information and automatically excludes abnormal heartbeats. Valid segmentation of the QRS complex plays a major role in further analysis since too restrictive segmentation might ignore fQRS patterns in the Q or S waves and the inclusion of artefacts or small oscillations might be confused with fragmentation. In previous work, this process has often been based on a single ECG lead, ignoring the complementary information available in multi-lead recordings. Second, the proposed machine learning algorithm for fQRS quantification is trained and evaluated in multi-center data, using different fQRS criteria comprising narrow and broad QRS complexes. Finally, to the best of our knowledge, this is the first evaluation of such an approach on non-sinus rhythm data, extending the applicability also to atrial fibrillation signals (AF). These models were tested and their performance is presented in comparison with visual fQRS detection.

4.2 Methods

The block diagram in Figure 4.1 illustrates the methodology proposed. The algorithm uses all leads for segmentation, but feature extraction and classification are applied to each lead independently. As in the Das criteria for region specific fQRS analysis, all leads, except aVR, were used in the analysis [38, 176]. Hence, the output presents 11 fQRS scores per patient's recording. In the remainder of this chapter, each lead will be referred to as *signal*, and the set of 11-lead ECG will be referred to as *recording*.

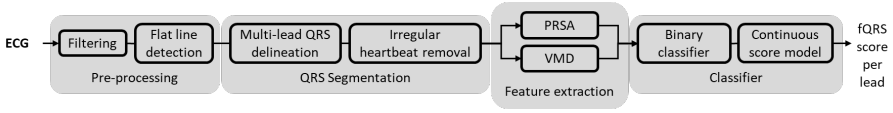

Figure 4.1: Block diagram of the fQRS scoring algorithm per lead.

4.2.1 Data

Two independent datasets were used in this study. The first dataset includes 723 digital ECG recordings (10 seconds 12-lead ECG, 250 Hz sampling rate) of patients before ICD implantation in prevention of sudden cardiac death at the University Hospitals Leuven (UZL) in Leuven, Belgium. For this analysis, only recordings in sinus rhythm and AF were included, corresponding to 673 recordings (616 sinus rhythm and 57 AF). The details of these patients are summarized in Table 4.1.

The presence of fQRS was annotated independently by 5 clinicians [175] (referred here as observers), who assigned a binary label to each of the 11 signals as fragmented

(i.e. 1) or non-fragmented (i.e. 0) by using the fQRS criteria for narrow and wide QRS complexes defined by Das et al. [36, 38]. Each signal in the database was assigned the sum of the independent labels given by the 5 observers, resulting in a joint label per lead with values ranging between 0 and 5 (Table 4.2). While labels 0 and 5 represent full agreement of the clinicians in the absence or presence of fQRS, respectively, the labels in-between represent inter-observer variability. The signals with full agreement were used as ground truth to train the algorithm, as these signals represent the highest level of certainty.

Table 4.1: Patient characterization for UZL and EU-CERT datasets. Expressed as number of ECG recordings and percentage between brackets, or mean ± standard deviation.

	UZL	**EU-CERT**
Demographics		
Age at implant (y)	62.4 ± 11.6	63.8 ± 10.9
Male gender (%)	570 (84.7)	1006 (79.8)
Ischemic cardiomyopathy (%)	446 (66.3)	869 (68.9)
Left ventricular ejection fraction (%)	32.3 ± 12.3	24.9 ± 6.2
Primary prevention (%)	381 (56.6)	1261 (100)
Sinus rhythm (%)	616 (91.5)	1102 (87.4)
Atrial fibrillation (%)	57 (8.5)	159 (12.6)
Comorbidity		
Diabetes mellitus (%)	119 (17.7)	342 (27.1)
Stroke (%)	67 (10.0)	132 (10.5)
ECG		
Wide QRS (%)	370 (55.0)	496 (39.3)
Heart rate (bpm)	64 ± 13	72 ± 15
Follow-up		
Duration (y)	3.7 ± 3.4	3 ± 2.3
Appropriate Shock (%)	222 (33.0)	165 (13.1)
Death (%)	186 (27.6)	230 (18.2)

The second dataset was provided by the multi-center EUropean Comparative Effectiveness Research to Assess the Use of Primary ProphylacTic Implantable Cardioverter Defibrillators (EU-CERT-ICD) project [203]. This project included a retrospective registry of 5111 patients who received a primary prevention ICD between 2002 and 2014 in one of the 14 participating centers [161]. The cohort presented is based on the detailed ECG analysis reported by Pelli et al., 2020 [139]. After further exclusion of rhythms other than sinus and AF, a total of 1560 ECGs were included in the current analysis. This selection contains 402 12-lead recordings from University Hospital Basel, Switzerland (10 seconds, 500Hz, least significant bit resolution of 4μV provided by C.S); 29 from Oulu University Hospital, Finland (8 seconds, 1000Hz,

2.5μV, by T.K, H.V.H); 551 from University Medical Center Utrecht, Netherlands (10 seconds, 500Hz, 4.88μV, by A.T); 277 from University Medical Center Goettingen, Germany (10 seconds, 500Hz, 4μV by T.F, M.Z); and 301 from the UZ Leuven, Belgium (10 seconds, 500Hz, 4.88μV, by B.V, R.W). Since the ECGs from UZ Leuven are also part of the UZL data, these were excluded from the EU-CERT-ICD database, reducing it to 1259 ECGs. The details of this dataset are described in Table 4.1.

As described by Pelli et al., 2020 [139], the presence of fQRS was annotated independently by two clinicians, who later discussed a final binary label for leads at which they previously disagreed. The narrow QRS complexes were annotated using the criteria defined by Haukilahti et al., 2016 [74], while wide QRS complexes were annotated following Das et al., 2008 [38]. The binary label 1 of EU-CERT-ICD indicates agreement on the presence of fQRS, and 0 agreement on the absence of fQRS (Table 4.2). Since this dataset does not contain uncertain labels, the complete set of labels are considered ground truth.

Table 4.2: Percentage of leads per label in the UZL and EU-CERT datasets, both for sinus and AF files.

		Percentage files per label in UZL					
		0	1	2	3	4	5
UZL	sinus	40.95	13.19	7.23	5.76	8.16	24.70
	AF	41.15	13.56	10.85	5.90	6.06	22.49
EU-CERT	sinus	76.60					23.40
	AF	77.70					22.30

The inter-observer variability between UZL and EU-CERT-ICD labels was assessed by comparing the annotations of the 301 recordings shared by both datasets. This comparison was performed by binarizing the UZL labels for different thresholds from 1 to 5, and deriving the Kappa score κ, for each of these scenarios. The highest agreement (κ = 0.46) was obtained when all observers agreed on the presence of fQRS. The relaxation of this threshold by including cases in which fewer observers agreed on fQRS in the positive class implied a linear decrease of κ. Overall, there was only moderate agreement when using the different clinical criteria to assess fQRS. The criteria for fQRS scoring used in the EU-CERT-project were stricter than the Das-criteria used in the UZL-dataset. This could be expected since all signs of fragmentation were scored in UZL, while in EU-CERT-ICD some notches in the QRS complex were not considered as fQRS unless they represented one of the specific patterns described by Haukilahti et al., 2016 [74].

4.2.2 Pre-processing

A two-step pre-processing was applied. First, the signals were band-pass filtered using Butterworth filters of orders 4 and 6, respectively from 0.5 to 70 Hz, to remove baseline wander and high-frequency noise while maintaining the potential fQRS information. Second, each signal was normalized by removing the mean and dividing by the standard deviation of all voltage values.

4.2.3 QRS segmentation

Segmentation is crucial to properly estimate the level of fragmentation within a QRS complex. While it must be wide enough to capture the information in the Q and S waves, it needs to be robust enough not to confuse these waves with oscillations and noise around the QRS complex.

The first step of a robust QRS segmentation is detecting the R-peaks in the ECG signal, which is followed by identification of the surrounding Q and S waves. In this work, R-peak detection was performed using the algorithm included in the R-DECO software [122]. Since this algorithm treats every lead independently, the multi-lead information is integrated in a consecutive post-processing stage. The starting point is the number of heartbeats n_R in the recording, which should be equal in all signals from a same recording. The value n_R is defined as the most commonly found number of automatically detected R-peaks in all signals of one recording. The aim is to have a set $\mathbf{R}$ of R-peaks with n_R annotations per signal l. If more than n_R R-peaks are detected in a signal, only those n_R complexes closest to the average location of each R_i peak are kept. Conversely, in a given signal l_i in which less than n_R R-peaks are automatically detected, a second search is performed to detect the missed heartbeats by performing a refined maximum search in a window of 40ms in the absolute values of the signal l_i around the average position of the missed R-peak. The pseudo-code for this step is summarized in Algorithm 1.

After having identified the set $\mathbf{R}$ with the same number n_R of R-peaks in all leads, the QRS complexes were segmented. In a review by Beraza et al., 2017 [17], the segmentation method proposed by Martinez et al., 2004 [116], and used in the ECGkit software tool [43], was identified as the most accurate. This method is based on wavelets and was used to detect the onset and offset of the QRS complex for each lead independently. Next, similar as for R-peak detection, the multi-lead information is integrated in a post-processing stage presented in Algorithm 2. Once the QRS complexes are automatically delineated for each signal, the Q and S waves are reallocated per heartbeat to the median of all signals. Thereupon, these locations are adjusted using the information contained in the rest of the heartbeats of the same signal exploiting the periodicity of the signal. Hence, a set of rules is performed

Algorithm 1 Multi-lead Rpeak detection

Input: Multi-lead ECG recording
Output: Set **R** with n_R Rpeaks per lead l_i

1: Detect Rpeaks in each absolute lead, R_{l_i}, using Rpeak detector by Moeyersons et al., 2019 [122].
2: Calculate n_R as the most commonly found number of Rpeaks detected in all the leads.
3: **for** the leads in which the size of $R_{l_i} > n_R$ **do**
4: Keep only those n_R closest to the average location of each R_i peak.
5: **end for**
6: **for** the leads in which the size of $R_{l_i} < n_R$ **do**
7: Find missing peaks in absolute signal in a window of 40ms around the average R_i.
8: **end for**

per signal. First, the QRS complexes are segmented and normalized to a maximum amplitude of 1. Next, a template is selected per signal, defined as the QRS complex with the highest cumulative correlation to the other QRS complexes within the lead. The rest of the QRS complexes are aligned to this template, and their Q and S locations are adjusted accordingly. Subsequently, a quality-check is performed to remove irregular and erroneously delineated heartbeats. A vector $\rho = [\rho_1, \rho_2, \ldots, \rho_{n_R}]$ is calculated per heartbeat, containing the normalized correlations between it and the remaining heartbeats in the given signal. This vector has n_R elements and compares the given QRS complex to each of the n_R heartbeats in the signal. If more than half of the entries in vector ρ are lower than a quality limit q, where $q \in [0, 1]$, the given heartbeat is removed from further signal analysis. While q can be user selected, it was set here to 0.85 based on visual inspection by an experienced clinician. An example of the main steps of the segmentation strategy for a signal with irregular heartbeats is illustrated in Figure 4.2. The QRS segmentation methodology is publicly available together with illustrating demos (https://github.com/avillago/multiLeadSegmentation).

4.2.4 Feature extraction

The features included in this work are based on those proposed by Goovaerts et al., 2018 [64]. The features are extracted independently for each lead, and all the QRS complexes selected in the QRS segmentation stage are considered for feature extraction. The characterization of each signal is based on features extracted from variational mode decomposition (VMD), phase-rectified signal averaging (PRSA) and the number of peaks in the QRS.

Algorithm 2 Multi-lead QRS segmentation

Input: Multi-lead ECG recording, set **R** with n_R Rpeaks per lead l_i, quality limit q
Output: Final Q,R,S locations

1: Detect Q and S locations in each lead using the algorithm by Martínez et al., 2004 [116].
2: Adjust Q and S locations to the median of each point between all leads.
3: **for** each lead **do**
4: Segment each heartbeat and normalize by the maximum.
5: Align all heartbeats using correlation.
6: Select template heartbeat as the one with highest cumulative correlation.
7: Re-align rest of the heartbeats to the template and extract their correlation index ρ_i, l.
8: Remove the heartbeats with $\rho_i < q$ for more than n_R elements.
9: **end for**

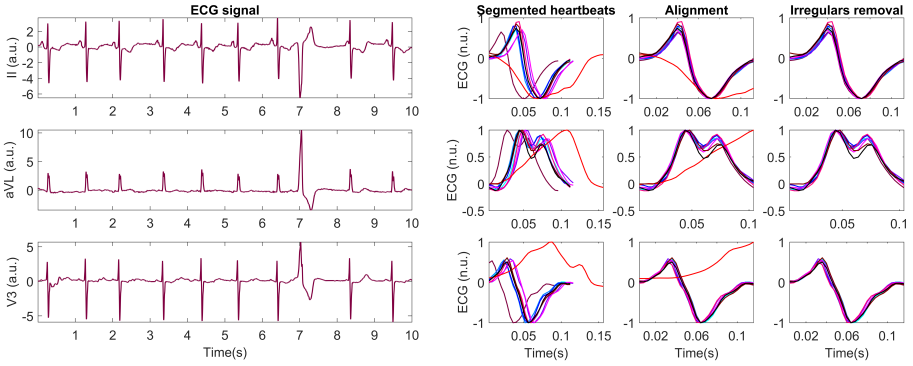

Figure 4.2: Example QRS segmentation for multi-lead signal with irregular heartbeats, showing the segmentation steps for leads II, aVL and V3. From left to right, the columns show the ECG signal, the segmented heartbeats (step 4 of Algorithm 2), the aligned heartbeats (step 5), and the final segmented heartbeats after removing irregular heartbeats.

VMD decomposes the signal into its main frequency components according to their level of contribution to the ECG waves. The number of modes was set to 5 based on previous ECG studies [64, 184], and these were sorted from lower to higher frequency. Since fQRS manifests itself as high-frequency notches in the QRS complex, it was shown that the central frequencies and morphology of the higher-frequency modes were different when comparing fragmented and non-fragmented signals [64]. An example of this decomposition for a fragmented and a non-fragmented signal can be seen in Figure 4.3. To capture these differences, the central frequencies and number of zero-crossings

of modes 3, 4 and 5 were used as features for classification, since these are the modes that show more differences between both classes.

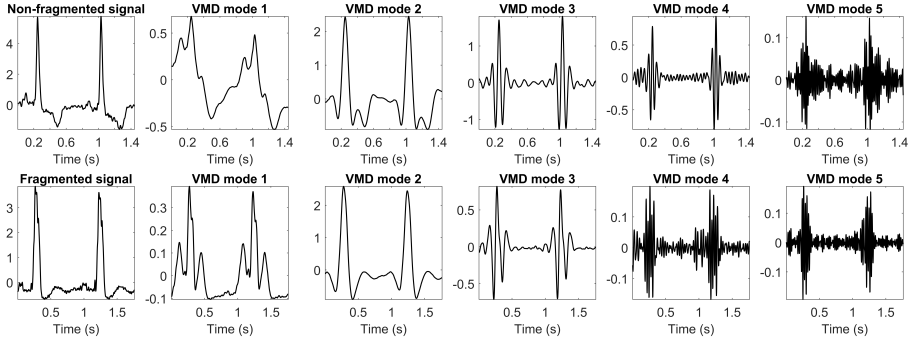

Figure 4.3: Examples of the VMD decomposition of non-fragmented (upper row) and fragmented (lower row) QRS complexes. Modes 1 to 5 are sorted in increasing frequency.

Additionally, PRSA was used to detect and quantify oscillations in the QRS complex, as disturbances of its expected trajectory. Applying a sliding window of length $2L$, consecutive overlapping segments of the QRS are averaged out to obtain an estimation of the global trend of the QRS complex, referred to as PRSA curve. This curve can be approximated by a linear fit. Non-fragmented signals present a clear slope in this approximation, characterizing the drastic amplitude change typical of a normal QRS complex. However, the QRS notches in fragmented signals flatten this line, since they disrupt the continuous trend of the QRS complex [66, 64]. Some examples of this PRSA methodology are shown in Figure 4.4. The first column shows the signal and some examples of the sliding window centers located along the QRS complex. The blue points correspond to samples in increasing segments of the signal, and the pink ones to descending points. The second column shows the PRSA curves extracted around each of these samples, with a length of 40ms. Those corresponding to descending points were inverted so they all follow an comparable trend. The last column shows in black the averaged PRSA curve on top of all the contributing curves. The features extracted from PRSA analysis are the mean derivative of the averaged PRSA curve and the slope and y-axis crossing of its linear fit approximation. Finally, the number of peaks in the QRS complex are also used as feature to describe the signal. The complete list of features is provided in Table 4.3.

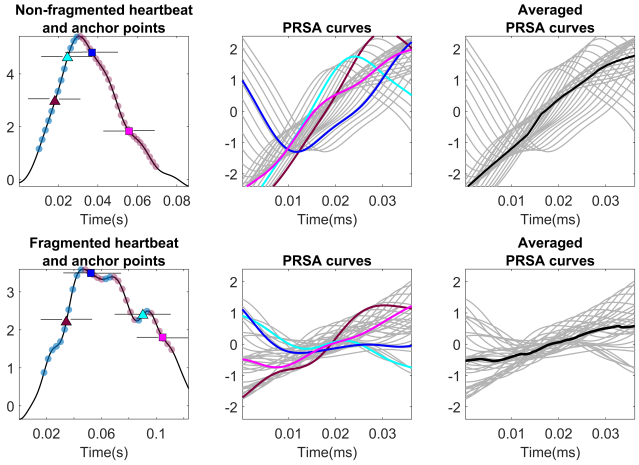

Figure 4.4: Examples of the PRSA procedure for the analysis of a non-fragmented and a fragmented QRS complex.

Table 4.3: List of features extracted from each lead.

VMD features	
Central frequencies	Mode 3
	Mode 4
	Mode 5
Number of zero crossings	Mode 3
	Mode 4
	Mode 5
PRSA features	
Mean derivative PRSA curve	
Slope linear fit	
Y-crossing linear fit	
ECG features	
Number peaks QRS	

4.2.5 Classifier

These 10 features extracted from each signal were used to train a Support Vector Machine (SVM) [33], a machine learning classifier, to discriminate between fragmented and non-fragmented signals. Each signal was introduced as an independent input, with no context of other leads [176]. The binary labels indicating the presence of fQRS, as described in Section 4.2.1, were used as an established baseline to train the classifier.

To account for possible non-linearities in the data, SVM with three different kernels were used to transform the feature space, namely the linear, polynomial and radial basis function (RBF). The parameters required to use them were tuned using Bayesian optimization. The SVM classifiers provide a binary output, which corresponds to fragmented or non-fragmented signals. To quantify fQRS, Platt scaling was used to convert the binary output of the model into a continuous score by fitting a logistic regression to the outputs[140]. By doing so, the model extracts the probability of a signal belonging to the fragmentation class.

4.2.6 Experiments

The different challenges of this study required independent training strategies and experiments to confirm their added value to the final algorithm. The first goal was to evaluate the impact of the new segmentation algorithm on the quantification of the fQRS. Secondly, we wanted to assess the effect of different fQRS definitions on the training and performance of the classifier. Finally, we evaluated the performance of the model in irregular signals form patients in atrial fibrillation (AF).

Experiment 1: The impact of the new segmentation algorithm

The impact of the proposed QRS segmentation algorithm was assessed by comparing its performance with the method proposed by Goovaerts et al., 2018[64], referred to as the *reference*. Both approaches were trained and tested on the same subset of the UZL dataset. First, training and evaluation was performed using only those signals in which all 5 observers agreed in the presence or absence of fQRS, i.e. labels 0 and 5. Patients were randomly split into training (80%) and test (20%), referred as *test1*. In order to reduce bias, this random division was performed 10 times, leading to training and evaluation with the reference and the proposed methodology in 10 different training-test sets. Next, the best model for each of these 10 repetitions was tested by adding the signals with labels from 1 to 4, both for the reference and proposed algorithm. This group of signals with annotations between 0 and 5 is referred to as *test2*. It allowed a comparison of the continuous score obtained from the algorithm with the certainty of fragmentation, represented by the agreement among the observers.

Experiment 2: The effect of different fQRS annotation criteria

The second experiment aimed at evaluating the effect of different fQRS annotation criteria in the multi-center data on the automated fQRS algorithm. By evaluating the algorithm in both UZL and EU-CERT-ICD datasets, its performance could be

assessed from two different perspectives, acknowledging the effect of the different fQRS definitions used. The final goal was to select a model that could be implemented in clinical practice by combining both approaches.

Three different training strategies were used:

1. Training exclusively on UZL. Here, the model selected in Experiment 1 was used, testing it in the corresponding UZL test signals and on the complete EU-CERT-ICD dataset.
2. Training exclusively on 80% of the data from EU-CERT-ICD. Given the lower percentage of fragmented leads in this dataset, a balanced set of signals was used for training the classifier. Therefore, only the same number of fragmented signals present in the training set were randomly selected from all the non-fragmented signals of the training patients. The other 20% of the EU-CERT-ICD patients were used for testing, as well as the complete UZL dataset.
3. The third strategy merged both datasets, using 80% of each dataset for training, while the remaining 20% of each was used as an independent test set. Similar to the previous strategies, out of these 80% of the patients, only the training leads labeled 0 or 5 were used in UZL, and a balanced set of leads was selected for EU-CERT-ICD.

A summary of the data assignment is shown in Figure 4.5. Given the random selections used to assign signals to the training set, each of these divisions was performed 10 times to reduce bias, and the set leading to the best results was kept in each case.

Experiment 3: Using the fQRS algorithm in irregular AF recordings

All experiments previously described were trained and evaluated on ECGs with regular sinus rhythm. Due to the morphology similarities of QRS complexes in sinus and AF signals, the introduced models should also apply to AF signals. In previous works, these signals were excluded from analysis due to their irregularities, since the short interval between heartbeats in AF might pose a problem to this type of algorithms. In the last experiment the best model selected from Experiment 2 was applied to the AF recordings of UZL and EU-CERT-ICD to assess the performance of the algorithm in ECGs with an irregular rhythm.

4.2.7 Performance metrics

The models were evaluated in test sets with different characteristics. Due to the lack of a “gold standard” for fQRS annotation, the signals scored 0 or 5 from UZL and

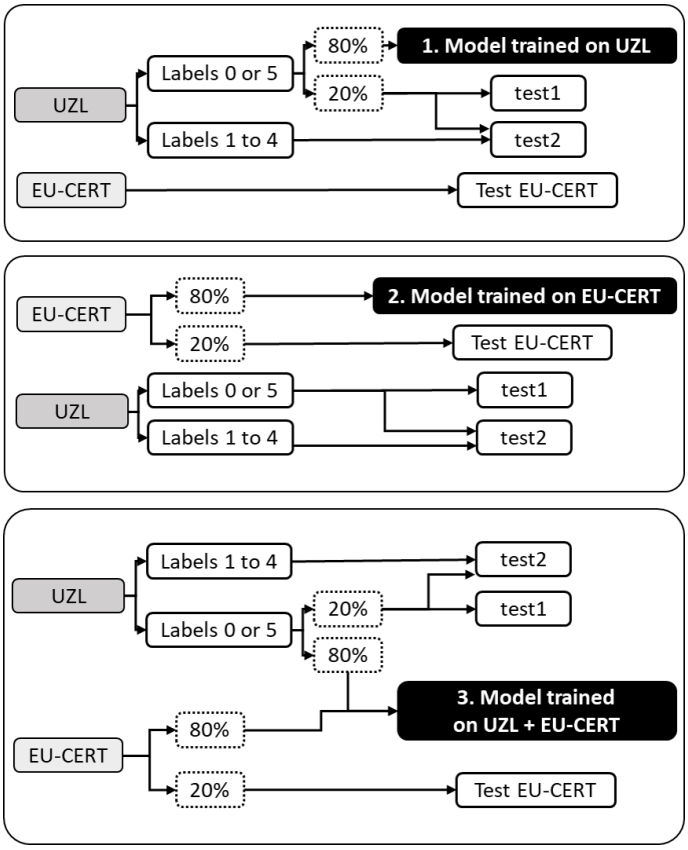

Figure 4.5: Block diagram describing the training strategies to evaluate the effect of multiple fQRS annotation criteria.

the binary labels from EU-CERT-ICD based on 2 observers are considered as the established baseline. Therefore, the results were evaluated using sensitivity, specificity, positive predictive value (Sens, Spec and PPV, respectively) and area under the curve (AUC) both for the receiver operator characteristic (ROC) and precision-recall (PR) curves. The latter is particularly relevant in the experiments that involve the EU-CERT-ICD dataset, given that its labels are highly imbalanced with fewer fragmented signals. In such a situation, the PR AUC provides more informative results than the ROC AUC [136]. These metrics were derived by binarizing the algorithm fQRS score using a 0.5 threshold. In the cases in which 10 iterations for each of the 3 kernels proposed were performed, the results of these 6 metrics are provided as the mean and the standard deviation. Additionally, Kappa scores (κ), reflecting the agreement between the algorithm scores and the manual labelling, are presented. The κ scores

were calculated several times for each experiment, exploring 100 different thresholds uniformly distributed between 0 and 1. The reported κ values correspond to the highest score obtained during these iterations. Lastly, the models were also evaluated on the UZL signals where observers did not agree. Since these signals involve uncertainty, the Pearson correlation coefficient between the labels and the continuous output of the models was calculated.

4.2.8 Analysis of feature importance

Once the best model from Experiment 2 was selected, the contribution of the features to the model output was explored performing a feature relevance analysis. This analysis was performed using SHapley Additive exPLanations (SHAP), which is a methodology based on game theory to explain the outputs of machine learning models and improve their interpretability [105, 106]. For every feature, the SHAP value quantifies the contribution of the future to the final prediction of the signal as fragmented or not fragmented. Positive SHAP values indicate that the feature positions the signal towards the fragmented class, and negative values towards the non-fragmented class.

4.3 Results

4.3.1 Experiment 1: The impact of the new segmentation algorithm

The sensitivity, specificity, PPV, ROC AUC and κ score results for the first experiment are shown in Table 4.4 and Figure 4.6. The first row of each block in Table 4.4 shows the average and standard deviation of 30 experiments: the 10 training-test combinations for each of the 3 kernels (linear, polynomial and RBF). The difference in results between the 3 kernels were minimal, as can be seen in Table A.1 of Supplementary Material, indicating that the feature space allowed a linear separation between both classes. Table A.1 presents the breakdown of the results for each of the kernels, as the mean and standard deviation of the 10 training-test combinations. Taking into account the similar results for all kernels, the models selected made use of the linear kernel for simplicity. The best training-test set was selected out of the 10 combinations and the results of the classifier using the linear kernel are shown in the *Selected model* row of each block in Table 4.4.

The newly proposed approach achieved better results than the reference for all metrics. The PR AUC values were 0.89 and 0.84 for the proposed and reference approaches,

respectively. This small difference can be explained by the high percentage of fragmented signals in the UZL test set making it almost balanced.

Table 4.4: Results of the proposed and reference approach on test1 of UZL including only signals with agreed labels between annotators. The first row of each block shows the average and standard deviation of each metric for all the iterations performed (10 training-test combinations for each of the 3 kernels). The second one shows the results of the selected models, corresponding to the best training-test combination for the linear kernel.

Algorithm		Sens	Spec	PPV	ROC AUC	κ
Proposed	Average + STD	0.72 ± 0.06	0.92 ± 0.01	0.86 ± 0.02	0.93 ± 0.02	0.68 ± 0.04
	Selected model	0.76	0.92	0.86	0.93	0.71
Reference	Average + STD	0.68 ± 0.02	0.89 ± 0.02	0.8 ± 0.02	0.88 ± 0.01	0.62 ± 0.03
	Selected model	0.68	0.89	0.79	0.88	0.62

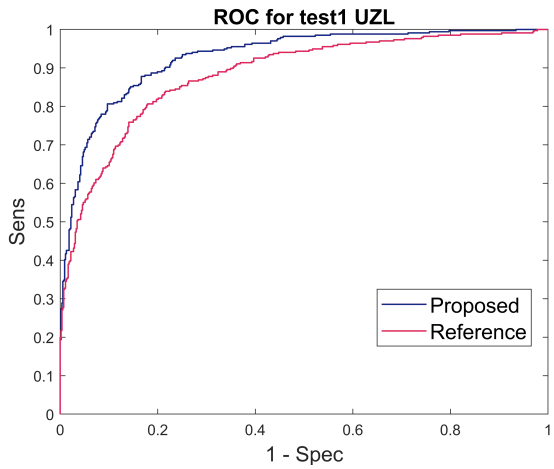

Figure 4.6: ROC curves for the proposed and reference approaches applied to test1 signals of UZL dataset, exclusively including signals without inter-observer variability.

Next, the continuous outputs were evaluated in the complete UZL-dataset with signals with labels ranging from 0 to 5. The boxplots in Figure 4.7 show the scores assigned by the proposed (purple) and reference (orange) model to the signals labeled 0 to 5. For both approaches the continuous score increased with the higher labels, illustrating the relation between the automated scores and the inter-observer variability. Pearson correlation coefficients were 0.50 for the proposed approach, and 0.45 for the reference (both $p < 0.001$). Additionally, in the proposed model, the scores assigned to the 0-labelled and 5-labelled signals were more skewed towards 0 and 1 values, respectively,

which is reflected in the average scores for each of these labels presented in Table 4.5. The scores assigned by each of the approaches to these two labels were statistically significantly different both for 0-labelled and 5-labelled. Therefore, we can conclude that compared to the reference algorithm, the proposed approach provides more consistent scores for the signals with inter-observer agreement.

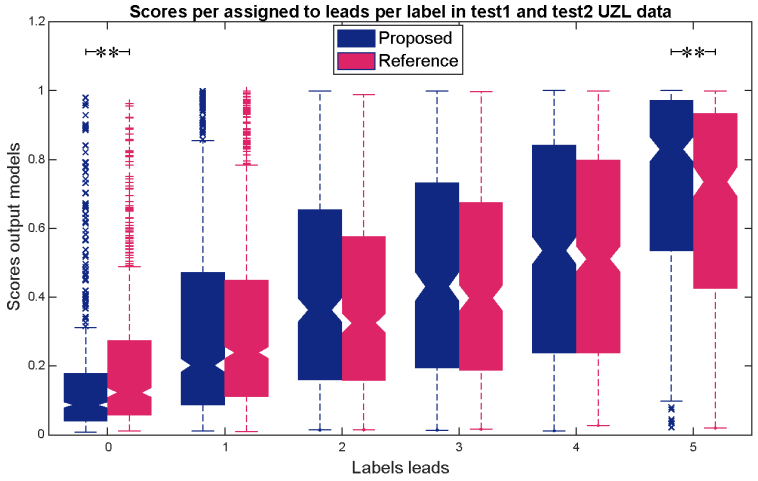

Figure 4.7: Relation between the output scores of the proposed and reference approaches (in black and grey) with the labels of test2 UZL signals, ranging from 0 to 5. Friedman test was applied to compare the results of both approaches, obtaining p-values below 0.005. These statistically significant differences are described by the sign ** in groups 0 and 5.

Table 4.5: Summary of the results of the proposed and reference approach in test2. The first two columns show the average ± standard deviation fQRS scores assigned to signals labelled as 0 and 5. The third and forth columns show the correlation coefficients between the fQRS scores and the ground truth labels from 0 to 5, both with p-values below .001.

	0	5	Pearson Corr.	p-value	95% CI
Proposed	0.16 ± 0.2	0.72 ± 0.29	0.50	$\ll .001$	[0.48, 0.53]
Reference	0.2 ± 0.21	0.65 ± 0.3	0.45	$\ll .001$	[0.42, 0.47]

4.3.2 Experiment 2: The effect of different fQRS annotation criteria

The results of experiment 2 are shown in Table 4.6 and Figure 4.8. For UZL test1 signals, the ROC and PR curves were very similar, whereas for EU-CERT-ICD, the PR AUC was lower than the ROC AUC. As expected, the best performance was obtained when the test set belonged to the same center as the data that was used to train the model. The trade-off achieved by the combined model was of interest since the performance for UZL was only mildly reduced whereas the performance in the EU-CERT-ICD test improved for sensitivity, and only decreased for specificity and PPV when compared to the EU-CERT-ICD trained approach. These variations might be attributed to the lower percentage of fragmented signals and stricter fQRS criteria in EU-CERT-ICD, which would increase the number of fragmented outputs in the combined approach due to the influence of UZL. This was confirmed by the κ scores since the best results on the EU-CERT-ICD data were achieved for thresholds of 0.5 for the model trained on the same dataset, and for thresholds greater than 0.8 for the models that included UZL data in the training stage. On the other hand, the thresholds that led to the best results in UZL were lower than 0.5 (0.42, 0.31 and 0.47 for training strategies 1, 2 and 3, respectively).

Table 4.6: Results of the 3 training strategies to evaluate the effect of multiple fQRS annotation criteria. From top to bottom, the table presents two blocks, the upper one shows the results for test1 UZL, and the one below for EU-CERT. Within each block, each row shows the results for the approach trained on UZL, on EU-CERT, and in both.

Training strategy	Sens	Spec	PPV	ROC AUC	PR AUC	κ
	test = UZL					
UZL	0.76 0.74 ± 0.02	0.92 0.92 ± 0.01	0.86 0.84 ± 0.02	0.93 0.92 ± 0.01	0.89 0.86 ± 0.02	0.71 0.69 ± 0.02
EU-CERT	0.59 0.57 ± 0.02	0.92 0.92 ± 0.00	0.84 0.84 ± 0.01	0.88 0.87 ± 0.00	0.83 0.83 ± 0.00	0.61 0.59 ± 0.02
UZL + EU-CERT	0.75 0.73 ± 0.02	0.90 0.90 ± 0.02	0.83 0.83 ± 0.01	0.91 0.90 ± 0.01	0.85 0.84 ± 0.02	0.68 0.66 ± 0.02
	test = EU-CERT					
UZL	0.90 0.90 ± 0.00	0.56 0.56 ± 0.01	0.37 0.37 ± 0.00	0.82 0.82 ± 0.00	0.53 0.53 ± 0.00	0.41 0.41 ± 0.00
EU-CERT	0.77 0.74 ± 0.03	0.78 0.77 ± 0.01	0.51 0.48 ± 0.02	0.85 0.83 ± 0.01	0.58 0.55± 0.02	0.48 0.46 ± 0.01
UZL + EU-CERT	0.88 0.86 ± 0.02	0.64 0.64 ± 0.01	0.42 0.42 ± 0.01	0.84 0.83 ± 0.01	0.57 0.55 ± 0.03	0.44 0.44 ± 0.02

Subsequently, the 3 training strategies were evaluated in UZL test2 signals. The continuous scores obtained for each of the 3 strategies were compared to the labels using Pearson correlation, obtaining coefficients of 0.5, 0.56 and 0.48 (all $p < 0.001$) for the UZL, EU-CERT-ICD and the combined model, respectively. Table 4.7 breaks down the results of the combined model by QRS duration both in test1 UZL and

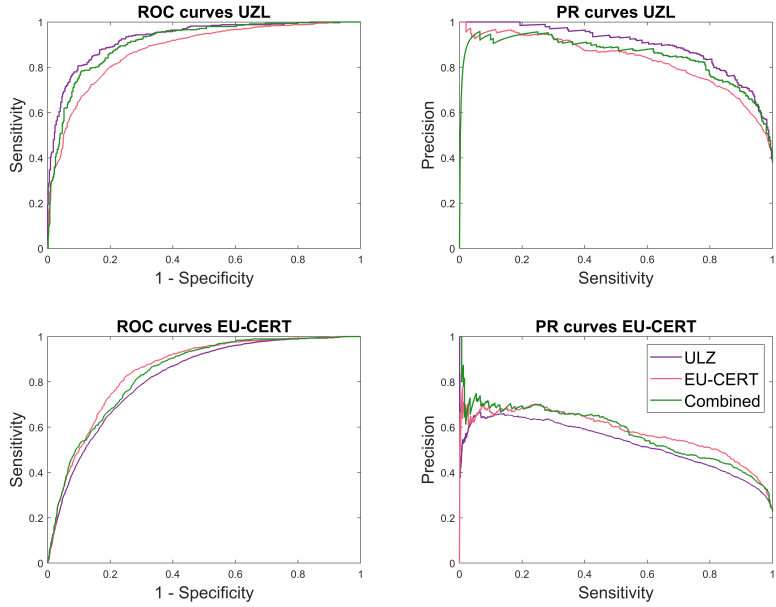

Figure 4.8: ROC and PR curves for test1 UZL and EU-CERT for the 3 training strategies considered.

EU-CERT-ICD data. There was a lower sensitivity when the algorithm was applied to narrow QRS complexes.

Table 4.7: Results for the combined model in test1 UZL and EU-CERT broken down by QRS duration.

	test = UZL						test = EU-CERT					
QRS duration	Sens	Spec	PPV	ROC AUC	PR AUC	κ	Sens	Spec	PPV	ROC AUC	PR AUC	κ
Narrow	0.69	0.91	0.86	0.91	0.87	0.68	0.86	0.68	0.53	0.85	0.67	0.49
Broad	0.83	0.90	0.80	0.92	0.84	0.73	0.90	0.62	0.33	0.84	0.47	0.44

4.3.3 Experiment 3: Using the fQRS algorithm in irregular AF recordings

Table 4.8 shows the results of applying the combined model proposed in experiment 2, to binary-labeled signals of UZL and EU-CERT-ICD with AF where the percentage of fQRS labelled signals is 22% and 33% respectively, compared to previously shown performance in regular sinus rhythm signals. As it can be seen, there was a sensitivity decrease in both datasets, accompanied by a slight increase in specificity and PPV. This

indicates that while the detected signal fragmentation was generally correct, some AF signals, that were considered to be fragmented on visual inspection, were missed by the model.

Table 4.8: Results of the combined model obtained from Experiment 2 in AF signals of both UZL and EU-CERT.

	test = UZL						test = EU-CERT					
Test rhythms	Sens	Spec	PPV	ROC AUC	PR AUC	κ	Sens	Spec	PPV	ROC AUC	PR AUC	κ
AF	0.63	0.96	0.89	0.90	0.86	0.63	0.74	0.73	0.43	0.82	0.54	0.43
sinus	0.75	0.90	0.83	0.91	0.85	0.68	0.88	0.64	0.42	0.84	0.57	0.44

4.3.4 Analysis of feature importance

Figure 4.9 shows the feature ranking according to their contribution to the final prediction of the combined model, and their influence in each of the test data. While SHAP values for a same feature may vary for different patients, in these figures the features are arranged according to their average SHAP value in order to classify their global importance. The color of the points corresponds to the actual class to which the signals belong: blue and purple points correspond to the fragmented and non-fragmented classes, respectively. As seen, the most important features are the number of peaks in the QRS complex, and the central frequencies of the highest VMD modes. However, this relevance ranking considers feature interaction, even though these interactions are not visible in Figure 4.9. While the three first most important features are common for both test data, there is a slight change of order from the 4th feature between both groups.

4.4 Discussion

We proposed and validated a new, fully automated machine learning approach for quantification of QRS fragmentation (fQRS) compared to the visual assessment of fQRS by clinical observers. The main contributions of our novel approach are:

- The definition of a new multi-lead algorithm for QRS segmentation, integrating information from different leads for a more robust QRS delineation. This method is publicly available at https://github.com/avillago/multiLeadSegmentation.
- Automatic removal of irregular heartbeats from the fQRS analysis process. This new QRS segmentation strategy includes a method to align the heartbeats in each lead, which allows deriving fQRS results only from normal and atrial premature heartbeats.

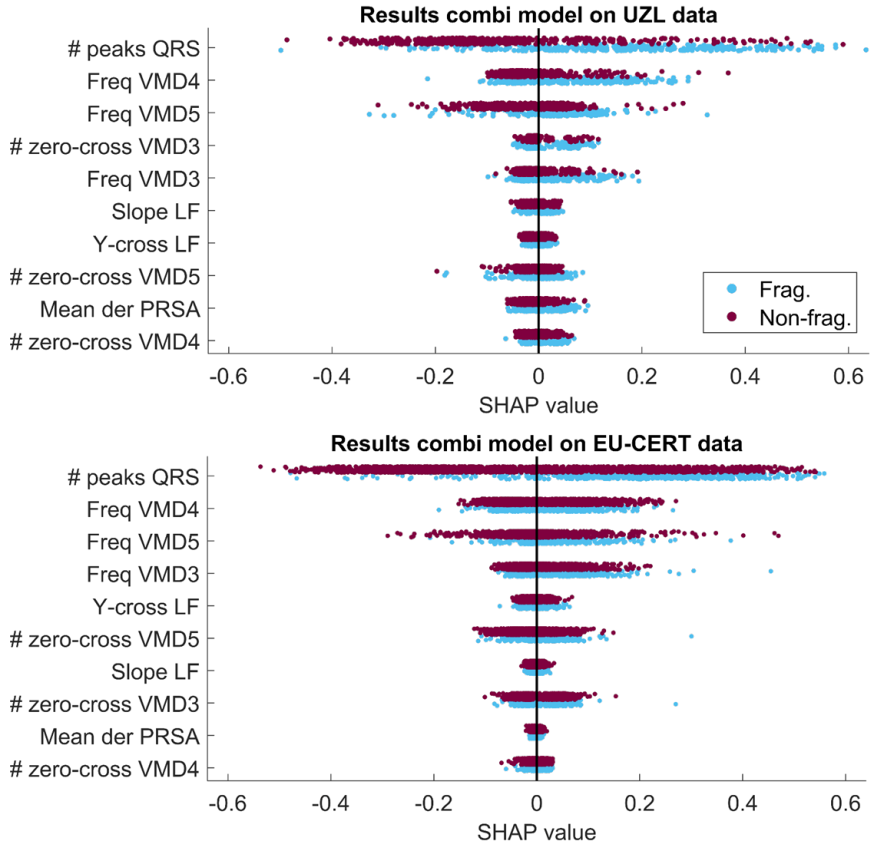

Figure 4.9: SHAP analysis of UZL and EU-CERT test signals for the combined model.

- Training and evaluation on a multi-center dataset, including ECG signals from 5 different European hospitals, allowing to assess the impact of inter-observer variability and the use of different fQRS criteria.
- The new algorithm was proven to be usable in both sinus rhythm and atrial fibrillation, as well as in narrow and wide QRS complexes.

The results show that the new QRS segmentation method improved the performance to detect visually confirmed fQRS, when compared to the earlier described reference method. The PPV and specificity indicate that the modified approach reduced the number of false positives. The sensitivity did improve as well. Further, the results indicate a better discrimination of the most evident fQRS cases compared to the reference, since the scores of signals in the UZL-dataset where all observers agreed (i.e.

labels 0, as non–fragmented, and 5 as fragmented) were more skewed to the extreme scores 0 and 1[64].

Currently, the only available comparator for our automated approach is the visual assessment of fQRS. This is not only prone to intra- and inter-observer variability[175], but different definitions of fragmentation have also been proposed for clinical practice. The impact of different fQRS criteria on the training and performance of our automated fQRS quantification method are of interest. Previous automated approaches were exclusively trained with a single definition of fQRS and involved fewer observers. In our study, the models trained exclusively on one dataset achieved low results when tested on the other one, due to their differences in fQRS criteria. The fQRS criteria used to score the EU-CERT-ICD dataset considered the magnitude and location of the fragmentation. Annotating only specific fQRS patterns and allowing some benign patterns to be labelled as non-fragmented, result in a more restrictive annotation and a lower prevalence of fQRS. The use of these criteria poses a challenge to the proposed algorithm, since the features used do not include information about the location of the notches detected. The VMD and PRSA features suggested by Goovaerts et al. exclusively inform about the presence of fQRS and are extracted for the entire QRS complex, with no information with regards to its location. This can be inferred from the results of the model trained and evaluated on EU-CERT-ICD data. This EU-CERT-oriented model detected several false positive classifications, as indicated by the lower specificity, PPV and the results of the PR-curves, when compared to the UZL-oriented model. The difference in fQRS criteria explains the opposing behavior in terms of sensitivity and specificity for both datasets. While the models in which UZL data was included in the training detect most of the fragmented signals of EU-CERT-ICD, this also resulted in a higher number of false positive classifications in the EU-CERT-ICD data. On the contrary, an exclusively EU-CERT-ICD-oriented approach achieved low values of sensitivity in the UZL data, but the combined strategy did not compromise the performance on the UZL data while offering a trade-off when testing on the EU-CERT-ICD data. This translated in higher sensitivity, detecting more of the actual fragmented signals with a paired increase of false positives which causes other metrics to decrease. However, this is an expected response in the EU-CERT data, given its reduced number of fQRS signal annotated due to its strict fQRS definition. Hence, we propose that the combined model would be the most generally applicable approach to clinical practice, combining the different views on fragmentation. While this approach combines different criteria presenting some strengths versus a single-database trained approach, its results are clearly influenced by the annotations used for the training. Therefore, this algorithm does not fully overcome the issue of inter-observer variability. However, in clinical practice the proposed approach could be used as a recommendation system for fQRS analysis. Those cases with more uncertain fragmentation scores, such as those ranging between 0.4 and 0.6, could require additional experienced visual inspection, while scores closer to 0 and 1 might be considered as recommended scores.

Figure 4.9 shows the relevance of the features in the output of this combined model. The main challenge of this evaluation is the presence of two different fQRS definitions, both for the training and evaluation stages. It can be seen that the feature relevance for UZL and EU-CERT data is slightly different from the fourth relevant feature. One reason for these differences is the large overlap between the fragmented and non-fragmented classes for the EU-CERT signals. This confirms the need of additional features for a better learning of the EU-CERT annotations, which could include characteristics such as the location of the notches in the QRS, as well as the number of notches. However, even the results for UZL do not show large SHAP values for the features related to PRSA. A reason for this could be the relevance of the QRS duration, which might lead to different results for narrow and wide QRS complexes due to the number of samples available (Figure 4.4).

The breakdown of results in terms of width of the QRS complex was not reported before for other automated approaches, despite its relevance in the fQRS analysis. Our results show a difference in performance depending on the QRS duration of the signals, especially in sensitivity and PPV in UZL data. The reduced sensitivity in narrow QRS complexes may indicate that the segmentation is more accurate in broad QRS complexes, but also that fragmentation is more evident in these signals. After observation of narrow QRS cases annotated as fQRS but receiving a low score from the algorithm, it was observed that many of these correspond to minimal fragmentation. The example in Figure 4.10 shows a minimal notch after the R-peak which is scored as fQRS when applying the criteria by Das et al., however the automated algorithm calculated a low score and considers this complex non-fragmented. These cases may be missed by the algorithm, particularly in narrow QRS complexes as there may not be enough samples capturing these abnormalities. These cases highlight the relevance of quantifying fQRS. Whether this minimal notching is clinically relevant is unknown and requires further investigation.

The second row of Figure 4.10 shows another example of a fragmented signal which received a low fQRS score by the automated algorithm. This is a specially challenging case, since fragmentation is only present in the end of the QRS complex. Additionally, this signal containing narrow QRS complexes and sampled at 250Hz becomes even more arduous, since the segmentation error is reduced to less than 10 samples. These particular types of heartbeats with notches in the Q and S waves may benefit from specific machine learning-based delineation algorithms, trained on representative examples.

Lastly, previous work never evaluated automated fQRS detection and quantification in AF recordings and algorithms were applied exclusively to regular sinus rhythm signals. In our study, the results obtained for these signals are in line with those of narrow sinus signals in UZL data. For EU-CERT-ICD, these results vary more. It should be noted that these results are affected by the low number of AF signals and the low prevalence of fQRS in AF signals.

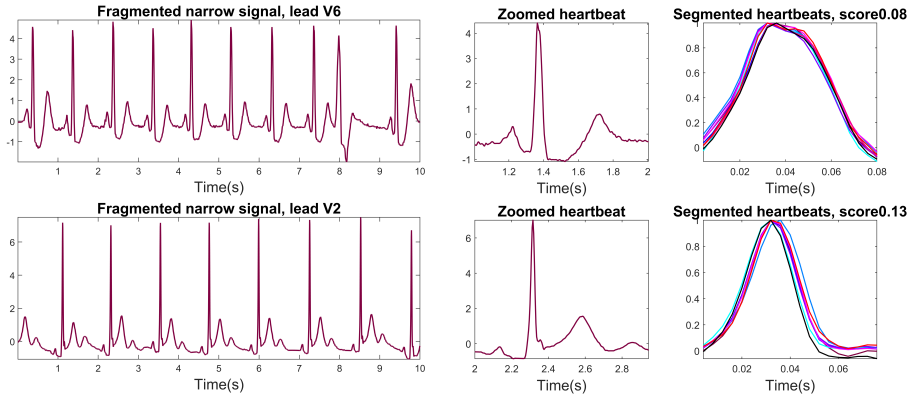

Figure 4.10: Example of two narrow QRS signals where the labels and the automatic score disagree (low algorithm score for fragmented signals). The top signal shows slight fQRS in the R-peaks, visible in the zoomed heartbeat and in the segmented heartbeats. The second presents fragmentation on the S wave, which lead to errors in QRS delineation and hence on scoring.

Limitations and future directions

In view of the known limitations of visual inspection of fQRS, the main limitation of our work is the training of our machine learning classifier using an imperfect standard. There is also no clinical consensus on the preferred fQRS definition. We tried to avoid this by using only the signals where all 5 observers agreed in the UZL-dataset thus avoiding inter-observer variability and by combining both criteria in the final model.

Since our goal was to technically validate the algorithm for signal analysis, we processed signals of separate ECG leads individually. In future work the obtained score per signal/lead needs to be summarized in a global score and a score per region/combination of leads (anterior (V1-5), lateral (I, aVL, V6) and inferior (II, III, aVF). We compared and validated the model to clinically used fQRS scoring. In the future, this continuous output of the fQRS algorithm could be used in clinical risk stratification by deriving a combined/regional score and correlating it to the clinical endpoints, e.g., severe ventricular arrhythmia or mortality. These regional scores have been shown to capture relevant information related to outcome. Vandenberk et al. showed that the presence of fragmentation in the inferior leads is related to the manifestation of arrhythmia; while fragmentation in the anterior leads is related to mortality [176]. However, both the regional and per-lead fQRS scores considered in that study are sensitive to inter-observer variability and would benefit from a standardized approach to identify fragmentation in order to derive relevant clinical conclusions. Machine learning applications could be useful to overcome the challenge of inter-

observer variability and deriving an optimal combination of scores to optimize risk stratification using fQRS. To this end, future work should explore using representation and unsupervised learning techniques to unveil hidden patterns relevant for this task.

4.5 Conclusion

This work proposed an automated machine learning method for fQRS detection and quantification. A novel, publicly available multi-lead QRS segmentation method was presented, combining the information of the multi-lead ECG and the different heartbeats in the signal. After segmentation QRS features were extracted, based on VMD and PRSA, they were used to train a machine learning classifier (Support Vector Machines) to discriminate between fragmented and non-fragmented ECG-traces. Data from 5 centers in which 2 different fQRS criteria were applied, were used in the training and evaluation of this new model. The combined model, which was trained using data from multiple centers and annotated with both fQRS criteria, showed to achieve comparable results to the specific models while making the model more flexible and more generally applicable to multi-center data, bringing it closer to clinical practice as a recommendation system. The algorithm was shown to be applicable to narrow and wide QRS complexes and regular and irregular rhythms, such as AF. These findings are crucial for the next steps to apply these machine learning models in clinical practice and assist clinicians in their diagnosis.

Chapter 5

Manifold learning for 12-lead ECG dimensionality reduction: A unified framework based on Laplacian Eigenmaps.

The analysis of multi-lead electrocardiographic (ECG) signals requires integrating the information derived from each lead to reach clinically relevant conclusions. Hence, this analysis could benefit from methods compacting the information in those leads into lower-dimensional representations (i.e., 2 or 3 dimensions instead of 12). In this work, we explore the value of Laplacian Eigenmaps (LE) to find structural and arrhythmogenic changes in the ECG.

We conceive a normal reference ECG space based on LE, calculated using signals of healthy subjects in sinus rhythm. Signals from new subjects can be mapped onto this reference space creating a loop per heartbeat that captures ECG abnormalities. A set of parameters, based on distance metrics and features to describe the shape of loops, are proposed to quantify the differences between subjects. This work was submitted for publication as:

> **Villa, A.**, Ingelaere, S., Jacobs, B., Vandenberk, B., Van Huffel, S., Willems, R. & Varon, C., A unified framework for multi-lead ECG characterization using Laplacian Eigenmaps. *Under review in IEEE Journal of Biomedical and Health informatics.*

5.1 Introduction

The electrocardiogram (ECG) is a graphical representation of the electrical activity of the heart and an important diagnostic tool in cardiology. The standard ECG consists of 12 leads containing information of different regions in the heart. The analysis of these multi-dimensional (i.e. multi-lead) signals can be challenging, since the integration of the information from the different leads is not straightforward. It is done by visually exploring the ECG taking into account the spatial information associated to the different leads [57].

Laplacian eigenmaps (LE) are a data-driven dimensionality reduction technique that allows identifying low-dimensional structures in high-dimensional data [16]. This algorithm captures non-linear characteristics of the data, while preserving locality-information. This is done by building a similarity graph, where the relations between the data points are emphasized. Using singular value decomposition (SVD), this new representation of the data is reduced to a lower-dimensional space that allows reconstructing the main characteristics of the data. This low-dimensional version of the data is referred to as a Laplacian Eigenmaps space. Based on the assumption that bio-electric signals are influenced by non-linear dynamics, several works have opted for non-linear approaches to explore these signals, instead of other linear methods like PCA or SVD [53, 187]. LE was first used for the analysis of dynamical changes in multi-lead electrogram signals in time by Erem et al., 2012 [54]. Each heartbeat is represented by a trajectory in the LE space, and authors quantified the changes between consecutive trajectories, which were then related to the effects of ischemia on the electrical properties of the heart. Erem et al. proposed an out-of-sample extension for the LE algorithm, which allowed to first build a LE space in a training phase, and later map other signals onto it [53].

One of the most relevant aspects of LEs is the enhancement of imperceptible changes in the signals. For example, Good et al. showed that LEs capture changes caused by ischemia more accurately than traditional signal-derived features [62, 63]. Moreover, these findings are highlighted by LE already during ventricular depolarization, while traditional metrics can only identify this in the ST segment representing ventricular repolarization, which is vulnerable to noise and artefacts [23, 31].

While those studies have explored dynamic changes within the data of a single subject, the comparison of multiple subjects is not straight-forward. The LE algorithm derives a space from the signals, and the out-of-sample extensions allow mapping back future instances of these signals to this original map [53, 62, 63]. Nevertheless, this LE space is signal-dependent, and hence the morphologies of the trajectories of different subjects in their corresponding spaces cannot be compared.

The goal of this work is to propose a new method that enables the characterization

of different subjects in a unified LE space. The main novelty is the definition of a reference LE space onto which new subjects can be mapped, therefore generalizing the out-of-sample extension properties of the algorithm. This space is created using ECG signals recorded during sinus rhythm from presumed healthy subjects with comparable demographics. Our hypothesis is that patterns in multi-lead ECG signals of patients with cardiac disease can be detected in this reference space. The ECGs of patients with cardiac disease are hypothesized to translate in different, more complex patterns than those of healthy subjects. These differences could be detected by analyzing the trajectories related to each type of subject, or even those of a particular subject at different moments in time. We propose a set of distance metrics to compare different trajectories, and other features to characterize their morphology.

This method was evaluated on the detection of structural and arrhythmogenic signs on the QRS complex of the ECG. First, the framework was applied to discriminate ECGs of healthy subjects from ECGs of patients with ischemic or non-ischemic cardiomyopathy. Additionally, we explored the utility of this approach for the discrimination between cardiomyopathy-patients with ventricular arrhythmia and patients without ventricular arrhythmia. The goal was to capture the differences between these groups in the reference space. This framework is publicly available in https://github.com/avillago/LEnormalSpace[1], where the codes to generate the reference space are available, together with examples of how to map new data onto it and how to derive the distance metrics.

5.2 Methods

5.2.1 Data

This work used ECG signals from three groups. The first population includes healthy subjects, the second gathers patients with ischemic and non-ischemic cardiomyopathy; and the last one includes cardiomyopathy-patients either with or without ventricular arrhythmia. The different populations were matched for age and female to male ratio to the smallest group, i.e the cardiomyopathy-patients with ventricular arrhythmia, more specifically, with a history of electrical storm (ES).

ES is defined as 3 or more appropriate interventions within 24 hours for life threatening ventricular arrhythmia by an implantable cardioverter-defibrillator (ICD) [50]. The ES patients were recruited from the database of patients with an ICD in the University Hospitals Leuven (UZL). This is a single-center registry containing information of all patients with ischemic heart disease (IHD) or dilated cardiomyopathy (DCM) implanted with an ICD between 01/01/1996 and 31/12/2018. The Ethical Committee

[1]The codes and the reference space will be made public upon publication.

of UZ Leuven approved retrospective analysis on this database. This database contains 12-lead ECG signals of 10 seconds, sampled at 250 or 500 Hz. Two different subsets of patients were extracted from this database according to their pathology and history of appropriate ICD interventions as surrogate for lethal ventricular arrhythmias. The first subset were patients with ES. Exclusion criteria were age at implantation ≤ 18 years, less than 6 months between ICD implantation and ES, ES induced by sepsis, and permanent ventricular pacing. In total, 32 ES patients were identified and matched for age, sex, underlying cardiomyopathy (including both IHD and DCM), and type of prevention with 32 ICD patients without any appropriate ICD intervention (non-ES). Baseline demographics are shown in Table 5.1, under the columns ES and non-ES. The ECG signals considered for the ES patients correspond to the last visit prior to the ES event, in which the signal is in sinus rhythm or atrial fibrillation (AF). The average time to ES of the selected signals is 4 months, with a standard deviation of 2.5 months.

In order to evaluate the framework on structural ECG changes, a subset of signals from ICD patients without any appropriate ICD intervention at the moment of database closure is selected. To exclude QRS-width as confounding factor, patients with a narrow QRS (QRS duration $\leq$ 120ms) were selected from the UZL registry. A total of 64 ECG signals on sinus rhythm or atrial fibrillation (AF) were selected, 32 from IHD patients and 32 from DCM patients. IHD patients have a documented history of myocardial ischemia, while DCM patients were patients with cardiac dysfunction in absence of significant coronary lesions. Baseline demographics of these subjects are presented in Table 5.1.

Table 5.1: Baseline demographics of the ICD database from UZ Leuven and the matched normal ECGs from PTB-XL.

mean (SD)	**H**	**IHD**	**DCM**	**ES**	**non-ES**
Age(years)	63.5 (7.9)	65.7 (7.8)	62.2 (5.7)	64.0 (9.5)	63.0 (9.9)
LVEF(%)	NA	35.4 (13.3)†	27.9 (6.8)	38.4 (13.6)	37.8 (11.2)
HR(bpm)	70 (12)	68 (16)	71 (16)	65 (10)	65 (8)
QRS(ms)	78 (14.8)	87.7 (12.7)*	89.4 (12.1)*	120.0 (32.5)*	111.4 (32.9)*
number (%)					
Male	29 (91)	28 (88)	22 (69)	29 (91)	29 (91)
IHD	NA	32 (100)	0 (0)	28 (88)	28 (88)
Sec.Prev.	NA	11 (34)	5 (16)	25 (78)	25 (78)

* Significant differences to the H group † Sig. Diff. between IHD and DCM
LVEF Left ventricular ejection fraction, **HR** Heart Rate, **QRS** QRS duration,
Sec. Prev. Secondary Prevention

The definition of the normal space required ECG signals from healthy subjects. These signals were extracted from the PTB-XL database [190], which contains 21837 12-lead ECG recordings of 10 seconds sampled at 500 Hz. These recordings include annotations according to the SCP-ECG standards about diagnostic and rhythm; as well as a small description of abnormalities in the signals [84]. For our aim, we considered

a subset of 865 signals from different subjects matching age and male to female ratio to the ES population, including only signals belonging to the normal class, sinus rhythms and without abnormalities. Out of these signals, 32 were semi-randomly selected to build the training space, keeping the ratio female/male and the average age comparable to the ES population. The details of this group are also shown in Table 5.1, under the column H. Albeit there is no detailed information about the comorbidities of the patients without abnormalities on their ECG, they were considered to be healthy subjects [190]. In the remaining of this chapter, the ECG signals of this group are referred to as normal ECGs. Despite the selection of patients with a QRS-width $\leq$ 120 ms for comparison, as it can be seen in Table 5.1, the QRS duration of the different groups was significantly different to that of the H group. Therefore, we performed an additional analysis to address this.

5.2.2 Preprocessing

The preprocessing of the signals comprises all the steps applied to the ECGs prior to the use of the LE algorithm. The different steps are illustrated in Figure 5.1.

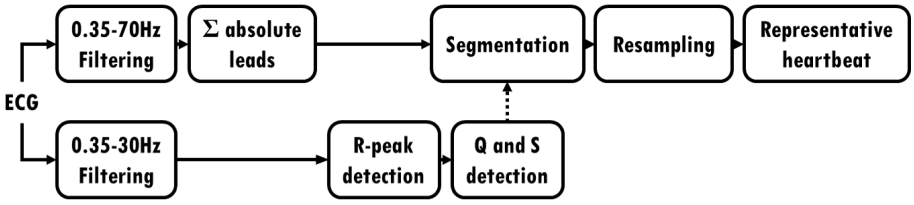

Figure 5.1: Preprocessing steps of the ECG signals prior to applying the LE algorithm.

First, the heartbeats were detected and delineated. In this work, we focus on changes in ventricular depolarization, hence only the QRS complex was considered in the analysis. To segment the QRS complexes, the signals were first bandpass-filtered between 0.35 and 30 Hz, to remove baseline wander and reduce the impact of high-frequency noise. Then the absolute value of each lead was calculated, and the addition of these 12 absolute leads was used to delineate the QRS complexes. Due to the differences in electrode placement for each of the leads, there are small delays in the timing of the ECG waves. Taking this into account, performing QRS delineation in the absolute added signal guarantees that the complete QRS complex is included in each of the leads. First, the R-peaks were identified using the algorithm proposed in the R-DECO software [122], and the onset and offset of the QRS complexes were detected using the ECGkit [116].

Even though high frequencies were removed for the QRS delineation process, the information of those frequency bands is relevant in our analysis. The presence of

high-frequency notches during the QRS complex can be associated to fQRS, which is an indicator of high risk of arrhythmia [176]. Therefore, the raw 12-lead ECG signals were bandpass-filtered between 0.35 and 70 Hz, and the location of the Q and the S waves detected before were imported to this signal, as indicated by the dotted arrow in Figure 5.1. Once the QRS complexes were segmented, they were resampled to a common length of N samples, here fixed to 200. After that, a representative heartbeat per signal was selected, as the one with the highest averaged normalized correlation with all the other QRS complexes of that recording.

5.2.3 Definition of the reference LE space

One of the main contributions of this work is the definition of a reference space onto which the ECGs of different patients can be mapped and compared in a unified framework. Given that the goal is to detect abnormalities in the QRS complexes, normal ECGs in sinus rhythm, were used to build this space. To exclude age- and gender- dependent changes, a subset of 32 normal ECGs were selected from PTB-XL, belonging to subjects that were matched based for their age and sex to the patient databases described in Section 5.2.1.

Using the signals of these subjects, a low-dimensional space was defined, where the underlying manifold structure of the data is captured. This means that the most relevant characteristics of the data are emphasized in this alternative space, while reducing the number of dimensions. The training and explicit mapping stages defined in [53] can be used here to calculate the space, and further used to map new data onto it.

The first step is to concatenate the representative heartbeats of each of the 32 normal ECGs. This data can be referred to as a set of points $X = \{x_i\}_{i=1}^{L} \in \mathbb{R}^{L\times d}$, with L the number of concatenated samples, i.e., 200 samples per heartbeat, 32 times; and d the number of dimensions (i.e., 12 channels) in the data. Both d and L can take different values according to the application. For the case of 12-lead ECG signals, the limiting factor in terms of computation is the number of samples L, since it determines the size of the matrix that needs to be stored in the memory and manipulated to derive the SVD. The structure of the data is shown in Figure 5.2. To emphasize the non-linear relations between the points X, a similarity matrix W is calculated using the pairwise kernel similarity between the points. Each entry of this matrix $W \in \mathbb{R}^{L\times L}$ is calculated as:

$$W_{ij} = \frac{-exp(||x_i - x_j||_2^2)}{\sigma^2}, \tag{5.1}$$

where i and j take values between $1,\ldots,L$; $W_{i,i} = 1$, $W_{i,j} = W_{j,i}$, and σ^2 is the radial basis function (RBF) kernel parameter. In matrix W, those pairs of points that are closer in the high-dimensional space receive a higher value $W_{i,j}$ than those that are very

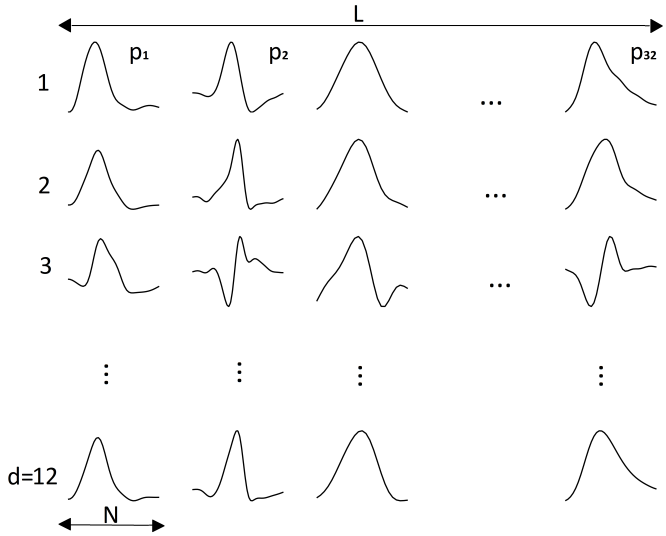

Figure 5.2: Structure of the data used to build the matrix required for the LE algorithm. The dimension L corresponds to N samples per reference heartbeat per subject (here N is fixed to 200 and the number of patients to 32). The dimension d is set to 12, the number of leads.

distant. The value of σ is set to the maximum $||x_i - x_j||_2^2$ value as suggested in [53]. Then the degree matrix D is computed, as a diagonal matrix in which each element is derived as $D_{i,i} = diag(\sum_j W_{i,j})$. This diagonal matrix captures the relevance of each point x_i in the manifold, since each entry $D_{i,i}$ is defined by the connectivity of the point x_i with the rest of the points in X.

These matrices allow the formulation of the singular value decomposition problem, defined as:

$$D^{-1}W = USV^T, \tag{5.2}$$

which enables the calculation of matrix V, where each column corresponds to one dimension of the Laplacian Eigenmaps space. These columns are sorted according to their associated singular values contained in S, from high to low values according to the amount of information they capture of the high-dimensional space. Since the first column is constant (all ones), the next ones are to be considered when constructing the LE space. In line with [16], the next three dimensions are used to define the space. These three dimensions capture >90% of the information contained in the data,

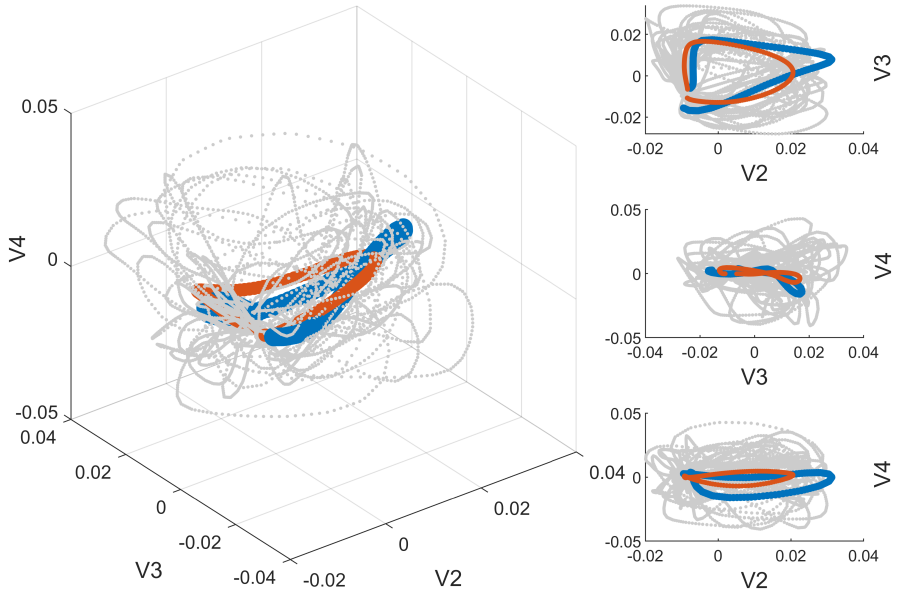

Figure 5.3: Trajectories of the ECGs used to build the reference space (grey). The template is shown in blue, and it corresponds to the trajectory that is closest to the average of all the training loops (in orange).

according to the singular values, and they allow intuitive visualization of the data [16]. The LE space is then defined by 3 vectors:

$$LE = [V_2, V_3, V_4]. \tag{5.3}$$

The LE space is characterized not only by its vectors V_2, V_3, V_4, but also by a template trajectory or loop. The representative heartbeats of each of the normal ECGs representing this LE space can be mapped onto it as explained in Section 5.2.4, creating a trajectory, as those observed in grey in Figure 5.3. All the loops fall within a manifold that could be represented by a template loop. This template is selected as the normal loop out of the 32 closer to the mean of all those used to create the space according to the euclidean distance. This mean is calculated point by point, obtaining an average loop of the same length as those of the different subjects. Figure 5.3 presents in grey the loops of the normal ECGs used to build the LE space, the template loop in blue, and the average used to specify the template, in orange.

5.2.4 Explicit mapping

As suggested in [53], the decomposition formulated in Equation 5.2 can be used to project multi-lead signals of new patients onto this manifold. In our case, the signal to be projected is the representative heartbeat of a new patient, extracted as explained in Section 5.2.2.

The explicit mapping function is defined as

$$f(y) = S^{\dagger} U^T D^{-1} g(y), \tag{5.4}$$

where

$$g(y) = [exp(-\frac{||y - x_1||_2^2}{\sigma^2})...exp(-\frac{||y - x_N||_2^2}{\sigma^2})]^T, \tag{5.5}$$

and $S^{\dagger}$ is the Moore-Penrose pseudo-inverse of matrix S. The 2nd, 3rd and 4th columns of the transposed output of $f(y)$ contain the low-dimensional representation of the test data $Y = \{y_i\}_{i=1}^{L} \in \mathbb{R}^{L \times d}$ [53].

The new heartbeats mapped onto this space also form a trajectory that can be compared to that one of the normal template heartbeat to identify differences. Figure 5.4 shows the examples of projected reference heartbeats for 10 ES and non-ES patients respectively. Additionally, the differentiable characteristics of this formulation allow mapping points back from the LE space to the original multi-lead signal.

5.2.5 Feature extraction

By creating the reference space and applying explicit mapping to the signals of interest, each of these is then characterized by a loop in the LE space. The goal is to identify specific characteristics of the patients that capture the complexity of their signals. Here, this complexity is represented by the differences between the trajectories of normal ECGs and those of the different patient groups.

Two types of features are proposed for this aim, namely distance metrics and features that intrinsically characterize the trajectories. They are summarized in Table 5.2.

Distance metrics

These features characterize differences as the distance between two trajectories. Following our previous work [85], three of the metrics proposed are further explored in this study. The first is the point-to-point distance which, given that all loops are

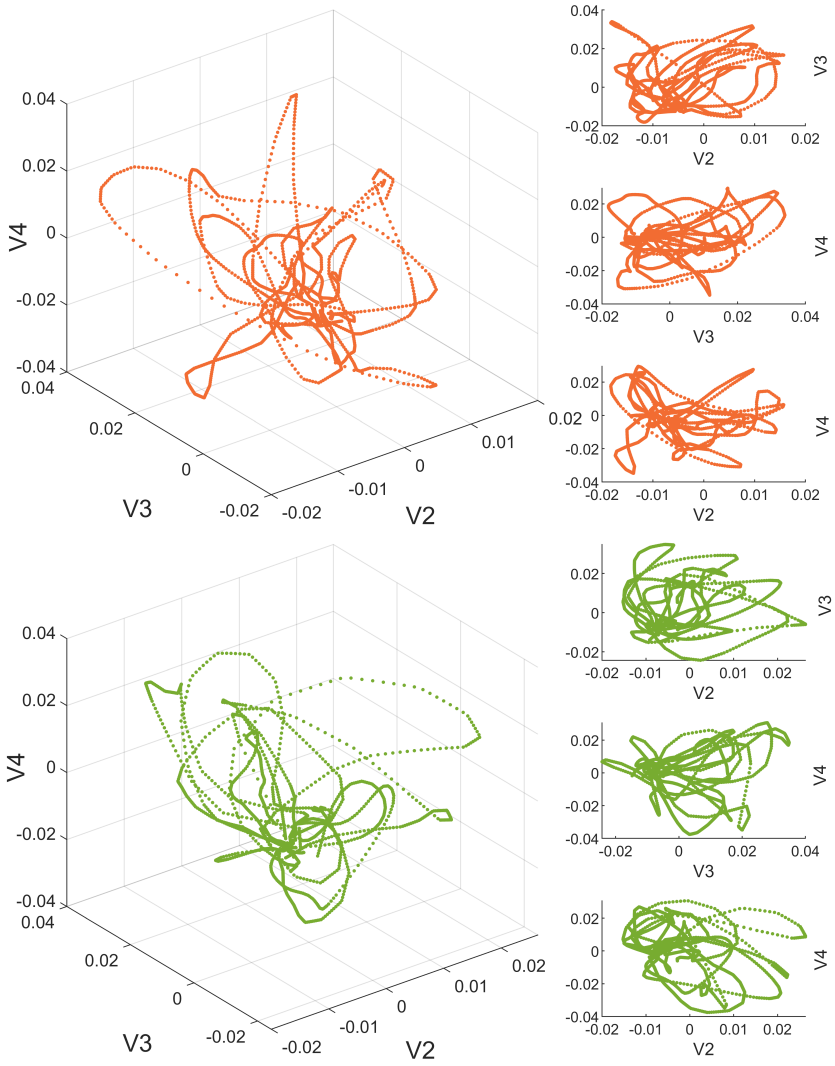

Figure 5.4: Examples of reference heartbeats explicitly mapped on the LE space for ES (orange) and non-ES patients (green).

Table 5.2: Features derived from the LE trajectories

Features	Notation
Distance metrics	
Point-to-point	d_{P2P}
Dynamic Time Warping	d_{DTW}
Hausdorff	d_{Haus}
Loop characterization	
Range V2	r_{V2}
Range V3	r_{V3}
Range V4	r_{V4}
SD V2	σ_{V2}
SD V3	σ_{V3}
SD V4	σ_{V4}
Mean curvature	$\overline{curv}$
SD Curvature	σ_{curv}

resampled to the same length, calculates the euclidean distance between each pair of points. The final distance is the average of all the distances d_i, with $i = 1, ..., L$. The second distance is based on Dynamic Time Warping (DTW) [127]. This allows a more flexible comparison of the loops by first finding their optimal match and then calculate the euclidean distance between the aligned time series. Similarly to the first metric, the final distance is calculated as the average of all the distances along the loop. The last metric is the Hausdorff distance, which is widely used to calculate the similarity between two sets of points [204]. This metric is calculated by finding, for each point x_i in trajectory A, its closest point in trajectory B, where trajectories A and B are the representative loops of two different patients A and B. The same is done for each point in B to trajectory A. The final Hausdorff distance is the maximum among all these shortest distances.

Loop characterization

The trajectories in the LE space can also be independently characterized, i.e. without comparing them to any other. Given the formulation of the LE problem and its relation with kernel PCA, the values of the trajectories in each dimension are related to their corresponding singular value and hence, to the amount of information that they capture. Therefore, the hypothesis here is that the simpler loops, corresponding to normal ECGs, should contain most of their information in $V2$. On the other hand, complex loops are those that differ from a normal sinus rhythm, and they shall distribute their information along the other dimensions. To capture this, the metrics proposed are the range and the standard deviation of the loops in each of the 3 LE dimensions. The range r_{Vi}

represents the difference between the maximum and minimum value of the points in each of the LE dimensions, and the standard deviation σ_{Vi} represents the dispersion of those values from the average. Additionally, the mean and the standard deviation of the curvature measured along the loop are also proposed to characterize the morphology of the loops mapped. The curvature for each point x_i of the loop is the inverse of the radius circumscribing points x_{i-1}, x_i and x_{i+1} [120]. Therefore, the average and the standard deviation of these values along the vector characterize the roundness of the loops in the space.

5.2.6 Experiments

In order to explore the potential of the proposed reference LE space to characterize different groups of patients, three experiments were performed. Firstly, we evaluated if the reference space is capturing normal properties of the signals. The second goal was to test the ability of the reference space to distinguish between ECGs of patients with different underlying pathologies, namely IHD and DCM. Finally, the last experiment evaluates the capability of the methodology to identify risk signs of arrhythmia. The reference space for all these experiments is the same, since all the groups of patients were matched on a population level.

Capturing characteristics of normal ECGs

The first step to assess the potential of this relative space was to verify if it was able to capture characteristics of normal ECGs. The hypothesis was that all normal sinus rhythm ECGs from healthy subjects of comparable age should present similar properties in this reference space.

To prove this, 10 separate groups of 32 normal ECGs, referred to as H_1,H_2,... H_{10}, were mapped onto the reference space. There was no repetition of subjects between the different groups, neither with those included in the creation of the reference space. For each of these groups, the features proposed were calculated and evaluated to check if all the 10 groups belonged to the same population.

Identifying cardiovascular diseases: differences between IHD and DCM

The second experiment aimed to verify if the LE reference space could distinguish between ECGs from healthy subjects and ECGs from patients of comparable age and gender with underlying cardiovascular diseases and a narrow QRS. To this end, the group of patients with no history of shocks from the UZL was mapped onto the reference space, including 32 IHD patients and 32 DCM patients.

The IHD patients considered have a history of myocardial infarction. Impaired coronary flow, mostly due to coronary atherosclerosis, can lead to hypoxia and subsequent cell death of cardiomyocytes in the region of the affected coronary artery [82]. This results in localized scar formation, which affects the conduction characteristics of the heart. Pathological Q-waves are a well-known ECG biomarker of previous infarction [42]. The second group of patients contains DCM patients. While in IHD the problem arises at the level of the coronaries, DCM primarily affects the heart muscle. There is a variety of causes that can underlie DCM, but they all lead to left or biventricular dilatation with systolic dysfunction [87]. Imaging is necessary for diagnosis, but some ECG biomarkers have been proposed to aid in the diagnostic work-up, for example low QRS voltages due to loss of cardiomyocytes, presence of left bundle branch block and T-wave inversion [55].

In this experiment, the goal was to evaluate if these two groups present different properties in the LE reference space, as well as if these are different than those of healthy subjects. The hypothesis is that IHD ECGs should be associated to the most complex loops in the LE space, in line with their morphology in the 12-lead format. On the contrary, DCM trajectories are hypothesized to differ from healthy subjects more subtle.

LE biomarkers for arrhythmia prediction: electrical storm

The last experiment explored the use of the LE reference space to identify patients at risk of electrical storm. For this aim, the 64 patients included in the ES study were mapped onto the reference space. Half of these patients had a history of ES and the other half did not. Both groups included IHD and DCM patients, in a similar proportion. Signals from ES patients were hypothesized to present possible signs associated with arrhythmia, such as fragmentation as explored in Chapter 3.

5.2.7 Statistical tests

The LE features were calculated for the different groups of patients. Significant differences between these groups were evaluated using the Kruskal-Wallis test ($p<0.05$) with Bonferroni correction for multiple comparisons. Correlation tests between features were evaluated using Pearson correlation.

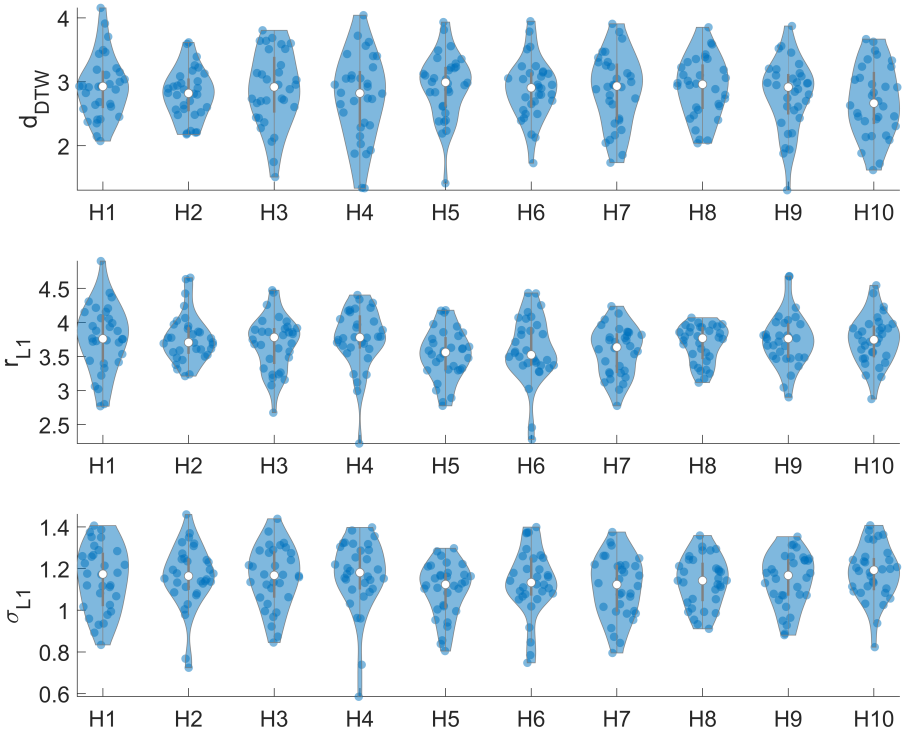

Figure 5.5: Violin plots for d_{DTW}, r_{V2} and σ_{V2} for the 10 groups of normal ECGs.

5.3 Results

5.3.1 Capturing characteristics of normal ECGs

The first experiment comparing 10 different groups of normal ECGs in the reference space did not find any significant difference between the 10 groups. Figure 5.5 shows the violin plots of the different normal ECG groups for some of these features. This suggests that the reference space was able to consistently characterize normal ECGs. Therefore, in the following experiments the groups of patients are compared to the group H_1 of normal signals, that can be considered as a representation of the PTB-XL database in the LE space. The average and standard deviation of the LE features for the H1 group are shown in Table 5.3.

Table 5.3: Results for H_1, IHD, DCM, ES and non-ES. The average and standard deviation values of the features are presented, together with the statistical comparisons per experiment (i.e. structural or arrhyhmogenic differences compared to H_1). Differences are significant for $p < 0.05$.

Features	H_1	IHD	DCM	ES	non-ES
d_{P2P}	3.4 ± 0.8	3.9 ± 0.8*,†	3.4 ± 0.8	4.3 ± 0.6*	3.9 ± 0.6
d_{Haus}	2.7 ± 0.5	3.0 ± 0.7	2.8 ± 0.6	3.2 ± 0.6*	2.9 ± 0.5
d_{DTW}	2.9 ± 0.5	3.5 ± 0.8*,†	3.0 ± 0.7	3.8 ± 0.7*	3.6 ± 0.6*
r_{V2}	3.8 ± 0.5	2.8 ± 0.8*	3.0 ± 0.7*	2.8 ± 0.8*	2.7 ± 0.8*
r_{V3}	4.0 ± 1.2	4.0 ± 0.9	4.5 ± 0.8	3.8 ± 1.2	3.7 ± 0.9
r_{V4}	4.7 ± 0.2	5.0 ± 1.1	4.8 ± 1.2	4.7 ± 1.1	4.6 ± 1.4
σ_{V2}	1.1 ± 0.6	0.7 ± 0.3*	0.8 ± 0.2*	0.7 ± 0.3*	0.7 ± 0.3*
σ_{V3}	1.1 ± 0.4	1.1 ± 0.3	1.2 ± 0.3	1.0 ± 0.3	1.0 ± 0.2
σ_{V4}	1.1 ± 0.1	1.2 ± 0.3	1.1 ± 0.3	1.2 ± 0.3	1.1 ± 0.3
$\overline{curv}$	27.0 ± 7.3	29.8 ± 10†	34.6 ± 7.7*	30.0 ± 14.6	29.9 ± 12.0
σ_{curv}	29.7 ± 9.3	28.9 ± 9.3	31.2 ± 9.6	33.5 ± 27.8	36.2 ± 30.8

* Significant differences to the H group † Sig. Diff. between IHD and DCM

5.3.2 Identifying cardiovascular diseases: differences between IHD and DCM

The features obtained for the IHD and DCM group are shown in Table 5.3. The symbol $*$ indicates significant differences (p-value < 0.05) between IHD or DCM patients and normal ECGs; and † between the IHD and DCM groups. The violin plots related to the features showing significant differences are presented in Figure 5.6, where these differences are indicated with $*$. As can be observed, the distance metrics identified significant differences between IHD and the other two groups, while the mean curvature $\overline{curv}$ of the trajectories allowed to distinguish between all the three groups. Additionally, features related to the first dimension of the LE space (r_{V2} and σ_{V2}) were significantly different between normal ECGs and the two groups of patients. However, this last distinction was also reflected in the significant differences between QRS duration of normal ECGs and patients shown in Table 5.1.

In view of the significant differences in QRS duration between the groups, the correlation between QRS duration and each feature was calculated, obtaining the highest values for the standard deviation of the values in the first LE dimension σ_{V2} (-0.27), and for the average curvature (-0.29). The upper plots of Figure 5.7 show the relation between these features and QRS duration. The shaded areas give a visual intuition of the separation that could be done between normal trajectories (blue points) and the patient groups (purple and yellow symbols).

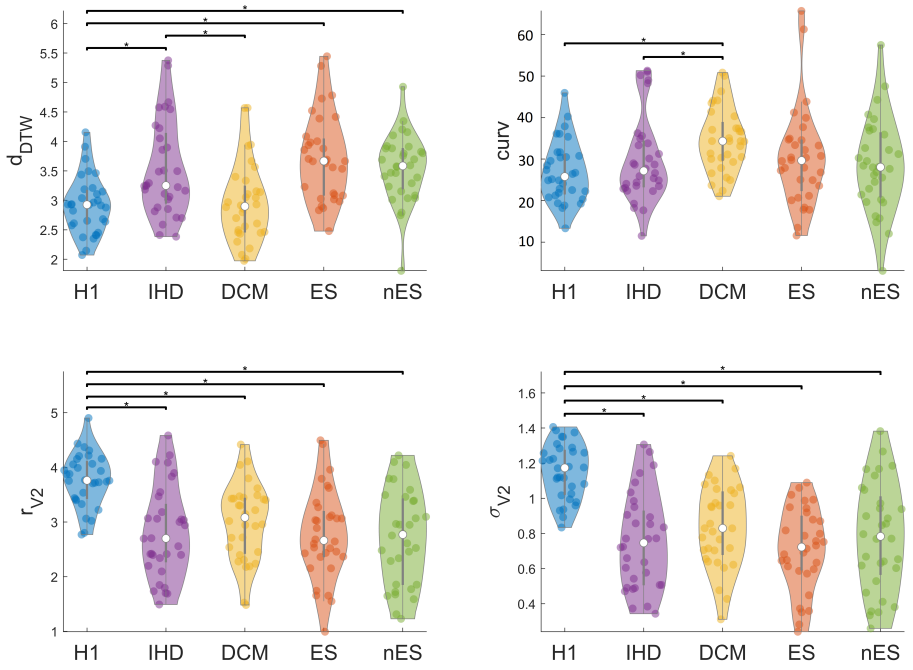

Figure 5.6: Violin plots for the features presenting significant differences for experiments 2 and 3. H1 presents the values for normal ECGs, compared to those of IHD and DCM patients, and to ES and non-ES. Significant differences are indicated as $*$ for p-value < 0.05.

5.3.3 LE biomarkers for arrhythmia prediction: electrical storm

The last experiment evaluated the use of the LE methodology for the characterization of patients developing ES. The average values of the features for both ES patients and their control group are shown in the last columns of Table 5.3. The LE features could not identify significant differences between the ES and non-ES group. Nevertheless, similarly to the results of the experiment described above, the space was able to distinguish between normal ECGs and these patients. The most relevant features for this distinction are the distance features and those related to V_2, both the range and the standard deviation. However, in this case the correlation of these features with QRS duration is stronger than in the previous experiment, since these patients have much wider QRS complexes than healthy subjects. The correlations with QRS duration for d_{P2P}, d_{DTW}, r_{V2} and σ_{V2} ranged between 0.44 and 0.57. Figure 5.7 shows the relation of d_{DTW} and r_{V2} with the QRS duration. Despite this weak correlation, the combination of QRS duration and these features contributes to differentiate between normal ECGs

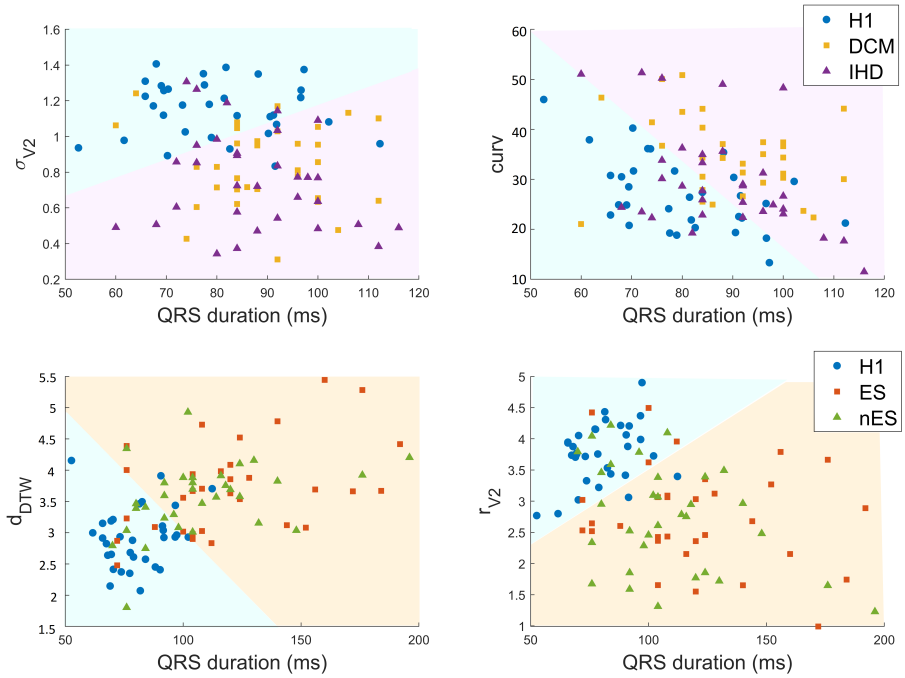

Figure 5.7: Relation between QRS duration and the features for the cases of higher correlation. In the top row, σ_{V2} and $\overline{curv}$ for Experiment 2 exploring DCM and IHD patients, and on the lower row d_{DTW} and r_{V2} for Experiment 3, exploring ES and non-ES patients.

(blue circles) and both ES and non-ES signals (orange and green signs), especially for QRS duration below 100ms.

5.4 Discussion

This study explored the potential of building a reference Laplacian Eigenmaps space based on normal multi-lead ECG signals. ECGs of other subjects can be projected onto this 3-dimensional space creating a trajectory per heartbeat. These trajectories capture the information of 12-lead ECGs and link it to the manifold information, representing the ECG in a lower-dimensional space.

The first experiment indicated that the LE reference space captures the similarities between normal ECGs. This was tested by projecting 10 groups of signals from the

PTB-XL database, and comparing the features obtained for each of these groups. No significant differences were found between them, which suggested that the method was able to detect that they all belonged to the same population. Thus, the LE features and the reference space proved to represent the characteristics of normal ECG signals.

Secondly, the methodology was applied to identify differences between three groups of ECG signals: those belonging to patients with IHD, DCM, and normal ECGs. The distances between the derived trajectories and the normal template were found to be significantly larger for IHD ECGs than for normal ECGs or those recorded from DCM patients. This indicates that the trajectories of IHD patients are more different from those derived from normal ECGs and DCM patients. While IHD and DCM are distinguishable using gadolinium enhanced cardiac MRI, where DCM often presents with patchy scars, this distinction is not yet possible using the QRS complexes of the ECG [118]. In IHD the electrical propagation over the heart is disturbed by scar tissue that affects the conduction through the ventricles, which leads to changes in the ECG [37, 42, 82]. Nevertheless, most of the well-known characteristics of IHD ECGs involve irregularities at the end of the QRS complex, which are difficult to detect, often presenting high false-positive rates [129]. Since we only considered QRS complexes in the analysis, the proposed methodology is able to identify differences in ventricular depolarization. A similar finding was reported in [63], where LE features were able to detect anomalies in acute ischemia ECG signals during the QRS complex, while traditional metrics required the analysis of longer segments including the T wave. Nevertheless, [62] and [63] study the evolution of signals over time, without comparing the loops of different subjects. Our methodology enables the analysis of different subjects in a unified framework, as well as the progress of these signals over time in a similar trend to previous works [62, 63]. This detection of signs of ischemia in the analysis of the QRS complex may suggest a big advantage of LE methods especially in noisy signals, since the low amplitude of the T-wave makes it more sensitive to artefacts [31, 23].

Additionally, the analysis of the mean curvature of the trajectories of the three groups showed significant intergroup differences, with higher values for DCM signals, followed by IHD and the lowest values obtained for normal ECGs. This metric can be associated with the complexity of the ECG signal: higher values of curvature describe more convex curves, while a regular QRS complex is expected to be characterized by a smooth rounded trajectory. Therefore, the most complex loops should be those of cardiomyopathy patients, and the simplest those related to normal ECGs. The discovery of significant differences in this aspect between healthy and DCM ECGs is nevertheless remarkable, since the characteristics of DCM ECGs are not evident by visual inspection [154]. Nevertheless, the presence of diseased cardiomyocytes and patchy scar in the heart of patients with DCM can be the cause of micro-fragmentation in the QRS complex [157], which could be captured by the LE algorithm.

These differences were also captured by features r_{V2} and σ_{V2}, which characterize the

trajectories in the first dimension of the LE space. This first dimension, due to the properties of the singular value decomposition, is the one that contains most of the information when the trajectories are simple, i.e. more similar to a normal regular sinus rhythm. More complex heartbeats, which may include fragmentation or conduction delay will additionally contain more information in the second and third dimension. This rationale has been considered in the decomposition of the ECG using SVD to characterize micro-fragmentation [77], and it can be applied to the Laplacian considered here. This intuition is shown in the results for r_{V2} and σ_{V2}, which are higher for normal ECGs than for DCM and IHD patients. While these features were slightly correlated to the duration of the QRS complex, as it can be seen in Figure 5.7, both aspects are complementary for the distinction between normal and abnormal ECGs.

Finally, the value of this methodology to detect features related to the future occurrence of arrhythmia, was tested. The goal was to detect differences between the ECG signals of patients who presented with electrical storm, and those who did not. However, none of the described features could detect differences between these two groups of patients. Currently no clear ECG biomarkers for risk prediction of this serious arrhythmic event are known. While the work presented on Chapter 3 linked the manifestation of ES to a trend of increasing QRS fragmentation, the exploration of a single visit for ES and non-ES patients could not identify significant differences. This database, compared to the one used in Chapter 3, was created using more restrictive inclusion criteria, reducing the number of signals. Further studies should explore comparable times between recordings, to reduce the impact of the time to event, which was variable and long in our series. Additionally, our dataset mixed patients with IHD and DCM, which have different LE trajectories. Therefore, future work in this field should try to study the ECGs of both populations separately to discard confounding effects, similarly to how it was done with other non-ECG biomarkers [9, 167].

Similarly to the IHD-DCM study, the ECGs of the patients included in the ES study were found to be significantly different from those of healthy subjects. Nevertheless, these findings were moderately correlated to the duration of the QRS, which was much larger for the group of patients than for the normal ECGs as reported in the literature [29]. Figure 5.8 shows the representative QRS complexes and the trajectories of the template loop (blue), its closest non-ES (green) and the farthest narrow ES (orange). The ES QRS complex in this case is fragmented in leads II and aVF, while the template and the non-ES signals are not. This could be the reason why the ES trajectory in the LE space contains more information in the third dimension, while the closest cases lay mostly in the horizontal plane (V2 and V3). This case enhances the relevance of the LE space to emphasize these differences, regardless of the QRS duration of the signals.

The aforementioned experiments confirmed the applicability of the reference LE space and the associated features for the analysis of signals of different subjects in a common framework. Nevertheless, some research questions remain open. Firstly, the current work exclusively explores changes in ventricular depolarization, since only the QRS

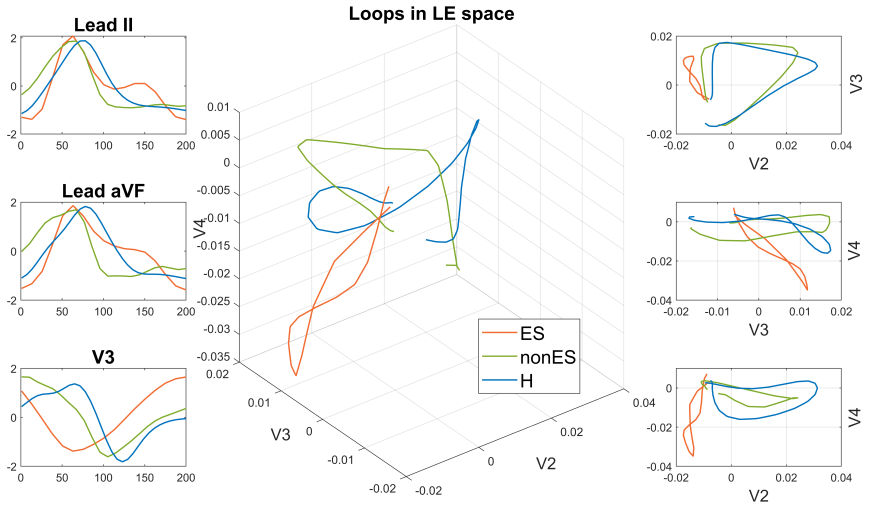

Figure 5.8: Example of 3 QRS complexes in the LE space and their corresponding ECGs. The ES signal (orange) corresponds to the farthest trajectory to the template, among the narrow QRS complexes. The green one is the closest narrow trajectory to the template (blue).

complex is included in the analysis. The inclusion of the T wave in the LE algorithm should be considered for further studies. Secondly, the proposed reference space was built using signals of subjects only with ages between 50 and 75, to match these to the age of the ICD population. The behavior of these maps may differ for signals related to different ages, sex and other clinical characteristics, which deserve further investigation. Additionally, the proposed approach could be applied to other signals, such as electrograms or electroencephalograms by defining a common reference space built with normal signals in this context.

When applied to ECG signals, this approach has similarities with the vectorcardiogram, which offers a 3-D representation of 12-lead ECGs [188]. While LE offer the added value of capturing non-linearities and perform a data-driven transformation, exploiting the information contained in the data, future work should explore its relation with VCG. Additionally, the current formulation assumes the presence of non-linearities in the data and it uses a kernel in the creation of the graph. However, neither this study nor the approaches proposed in [62, 63] have explored the use of a linear formulation for this aim. Future work should assess the impact of using either linear or non-linear embeddings, and propose guidelines about when to use each of them according to the type of data under investigation.

5.5 Conclusion

This work proposes a novel approach for the analysis of high-dimensional signals in an alternative space derived using the Laplacian Eigenmaps. This data-driven methodology allows reducing the dimensionality of the space spanned by multi-lead signals while emphasizing abnormalities and differences between subjects.

We evaluated this methodology in the context of 12-lead ECG signals, in the analysis of QRS complexes of signals recorded during sinus rhythm and AF. The methodology was able to capture characteristics of normal ECGs, and to distinguish these from those of patients with ischemic heart disease and patients with dilated cardiomyopathy implanted with a defibrillator in the prevention of sudden death by cardiac arrhythmia. In particular, the identification of differences between normal and DCM ECGs highlights the possible relevance of the proposed approach to identify differences not directly visible in the ECG. This novel methodology opens the path for future analysis of multi-lead ECGs and other high-dimensional signals in a common low-dimensional space, providing a new field of exploration in which subtle ECG abnormalities can be emphasized.

Chapter 6

Vectorcardiogram analysis for patient characterization

The previous chapter proposed the use of a novel approach based on Laplacian Eigenmaps to reduce the dimensionality of 12-lead ECG signals and enhance the discrimination of abnormalities caused by different pathologies. Nevertheless, the idea of capturing the electrical cardiac cycle in a low-dimensional space taking into account spatial information is not new in clinical practice. Since the 1950s, vectorcardiography (VCG) is used to quantify and visualize the magnitude and direction of the electrical activity of the heart in 3 dimensions. Additionally, the VCG can be derived from 12-lead ECG, and therefore, it can also be interpreted as a reduced-dimensionality representation of 12-lead ECG.

With this in mind, this chapter compares the LE methodology with the traditional VCG, and it highlights the added value of LE for the identification of patients with structural and arrhythmogenic abnormalities. Moreover, novel VCG features are proposed based on [24], which proved to benefit the characterization of 12-lead ECG into a reduced set of dimensions.

6.1 Introduction

The vectorcardiogram (VCG) is a well-known tool to capture the electrical activity of the heart in three dimensions. The VCG provides a graphical representation of the direction and magnitude of the heart vector over time, resulting from the addition of the electrical dipoles generated during the depolarization and repolarization of cardiac tissue [25]. It was first recorded in the 1950s, using three orthonormal leads, which allowed to build a 3D space conform with the 3D structure of the body [25]. This recording set-up, allowed representing the electrical activity of the heart over time, adding spatial information to ECGs, which were already in use. Different lead systems have been proposed to record this signal [111]. Figure 6.1 shows the electrode positions for the Frank VCG and their relation with the anatomical planes of the body.

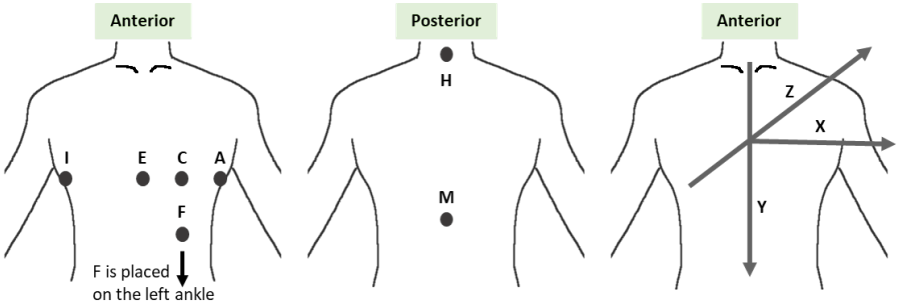

Figure 6.1: Electrode positions for Frank VCG recording. At the right, AHA axis directions used as reference coordinate system [111].

The trajectory of the cardiac vector in this 3D space creates a set of loops for each heartbeat, corresponding to the main waveforms of the ECG. Figure 6.2 shows an example of a VCG. The largest loop, in blue, corresponds to the QRS complex, while the two small loops capture the information in the P and the T-waves (red and green, respectively).

The analysis of the morphology, direction and relation between these VCG loops has been widely explored for different patient populations [112]. The most relevant findings of VCG are related to the study of myocardial infarction. Several characteristics have been proven to discriminate between healthy and infarcted patients [32, 146]; to assist the localization of ischemia [162]; and for risk stratification of all-cause mortality, sustained VT, or arrhythmic events [202].

Due to the need for different electrode positions and recording systems from that of the 12-lead ECG, and to the diagnostic advances of the standard 12-lead ECG, the use of VCG was reduced over the 1970s [112]. However, by the beginning of 1990s, there

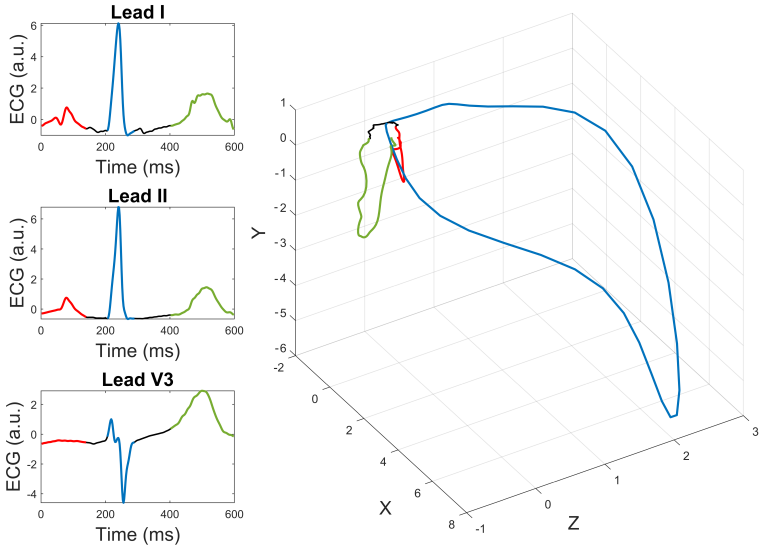

Figure 6.2: Example of VCG loops in 3D space. At the left, only 3 leads out of the 12 are shown for simplicity of the Figure. The waveforms present in the ECG can be seen as different loops in the VCG.

was an increased interest in VCG research, due to the formulation of methods to derive VCG from 12-lead ECG [188]. Such formulations removed the need of a separate recording set-up and added new insights into the ECG.

Several approaches have been proposed to derive the VCG from a 12-lead ECG signal, and they all are based on exploiting the redundancy of the 12 leads to reduce dimensionality and enable a simple visualization of the heart dipole [49, 95, 112, 188]. The most commonly used methods are the Inverse Dower Transform (IDT) [49] and the Kors transform [95], which consist of applying a fixed linear transformation to the signals, combining the information of leads I, II, and V1 to V6 [188]. The coefficients of the IDT method were obtained based on a mathematical model of the torso [49], while those of the Kors transform were derived by comparing recorded and calculated VCGs by means of regression analysis and minimizing the reconstruction error.

In this work, the Kors transform is used, since several studies suggest this approach as the one reconstructing Frank's VCG most accurately [86, 188]. Additionally, this method has been repeatedly reported to obtain a lower error when deriving the QRS-T angle, which is a commonly used VCG feature for arrhythmia prediction [34, 153]. The transformation functions to obtain the orthonormal VCG leads X, Y and Z in the Kors approach are:

$$\begin{aligned} X = &\ 0.38I - 0.07II - 0.13V1 + 0.05V2 - 0.01V3 + 0.14V4 + 0.06V5 - 0.54V6 \\ Y = &-0.07I + 0.93II + 0.06V1 - 0.02V2 - 0.05V3 + 0.06V4 - 0.17V5 + 0.13V6 \\ Z = &\ 0.11I - 0.23II - 0.43V1 - 0.06V2 - 0.14V3 - 0.20V4 - 0.11V5 + 0.31V6. \end{aligned} \tag{6.1}$$

The goal of this study is to explore the ability of VCG-based features to capture differences between the patient populations considered in the previous chapter. To this aim, a set of VCG features are extracted from the same groups of patients previously described, to compare the differences identified with those captured by the LE features.

First, a summary of state-of-the-art VCG features explored in studies including IHD and DCM patients, as well as for arrhythmia prediction is provided. Then the methodology proposed is described, including a mix of state-of-the-art features and novel tools developed during this thesis and formalized in [24]. A similar set of experiments to the one proposed in the previous chapter was performed, exploring these VCG features for the characterization of normal ECG signals; those of patients with structural abnormalities, namely IHD and DCM, and for patients with and without history of electrical storm (ES). We evaluate the findings of the VCG and the LE approach as a way to compare both ECG dimensionality reduction methods. Finally, some conclusions and future directions are presented.

6.2 State-of-the-art: VCG for IHD, DCM and arrhythmia prediction

6.2.1 VCG in IHD and DCM patients

One of the most studied conditions using VCG is cardiac ischemia and myocardial infarction. The presence of scar tissue in the ventricles causes deviations from the normal spread of the electrical activity over the heart, which are captured by the VCG in the form of abnormal loop trajectories [59]. In a study by Correa et al., the QRS loops of healthy subjects were compared to those of ischemic patients [32]. They showed that both groups were significantly different by characterizing the morphology of the loops. Normal loops were bigger in terms of volume and planar surface, and presented a more regular trajectory than those of ischemic patients. Other work has also analyzed the T-loop for this aim, showing that it can capture changes in ventricular repolarization, which have been linked to the level of ischemia [147]. Other relevant features are the angle between ventricular depolarization and repolarization (QRS-T

angle), which has been reported to be smaller for IHD patients than for healthy subjects [91], and repolarization lability, which is associated to higher risk of cardiac events [73, 113, 146].

VCGs of DCM patients can be expected to present some similarities to those of ischemic patients, though the abnormalities associated to DCM are expected to be more subtle [87]. To date imaging with for instance gadolinium enhanced cardiac MRI can aid to differenciate IHD from DCM, by the presence of patchy fibrosis in the latter, but there are no accurate ECG or VCG markers associated with DCM [118]. In one of the few studies on DCM and VCG interstitial fibrosis was induced in an artificial heart model in different regions of the ventricles and its impact on the VCG was analyzed [59]. The results showed that the QRS vector changed with the level of fibrosis, that was defined by the number of simulated small regions of scar [59]. These results confirm that larger scars and bigger ischemic regions lead to more abnormalities on the VCG. Hence, IHD VCGs are expected to be more different from normal VCGs, than signals from DCM patients.

6.2.2 VCG and arrhythmia prediction

The VCG has been largely explored for prediction of mortality and cardiac events. The most common VCG indicator of increasing risk of mortality is the QRS-T angle, both for the general and diseased population, and for several age groups [7, 90, 94, 194]. A wider QRS-T angle, which quantifies the angle between ventricular depolarization and repolarization, has been associated with higher risk of SCD both for IHD and DCM patients [61, 151].

Other features involve the morphology of the T-wave [103], and especially its beat-to-beat variability [73, 169]. While the latter has a well-known prognostic value, its analysis is not recommended in 10-second segments due to their limited duration.

Some studies have explored the relation of the QRS loop morphology and orientation with the risk of ventricular arrhythmias. Tereshchenko et al. identified a set of conduction characteristics associated with ventricular tachycardia that could be captured by the VCG, such as slower ventricular conduction, flatter QRS-loops and QRS folding [155, 170]. For a comparable analysis to the one suggested in our LE approach, our study explored this type of features, since they are derived from the QRS-loop.

6.3 Methods

The goal of this study is to compare the characterization potential of VCG and the LE approach proposed in Chapter 5. Accordingly, the same datasets described in Section

5.2.1 are used in this study, including the same subsets of signals. The following sections describe the methodology used to extract features from these signals and the tests performed.

6.3.1 Preprocessing

This section describes the steps prior to the feature extraction of the VCG loops. This process is divided in two big blocks, namely a first one to select the template heartbeat of the signal and the next one to delineate the VCG loops for further analysis. The sequence of steps is illustrated in Figure 6.3.

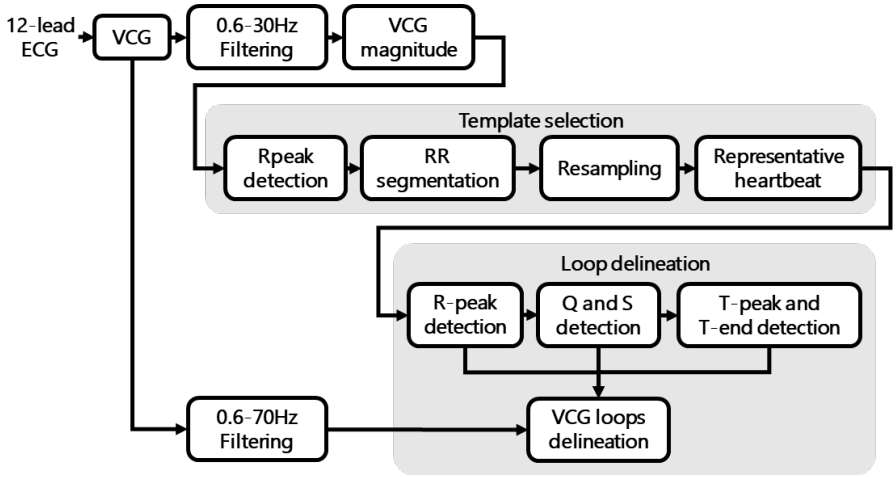

Figure 6.3: Preprocessing steps prior to the VCG feature extraction.

First, the VCG is derived from the 12-lead ECG using the Kors transform [95]. The VCG is then bandpass-filtered between 0.6 and 30 Hz for QRS delineation, so high frequencies do not mask the limits of the ECG waves. Subsequently, the magnitude of the filtered VCG is calculated using the Euclidean norm.

The magnitude of the filtered VCG is then used to select the template heartbeat, which similarly to previous chapters, is selected as the one with the highest correlation to the other heartbeats in the signal. First, the R-peaks are detected using the R-DECO software [122]. While previous chapters only considered the QRS complex for analysis, this work also includes the T-wave for the calculation of the QRS-T angle [135, 163]. Therefore, after detecting the R-peaks, each heartbeat is segmented using an asymmetric window around the R-peak. The last third of the previous RR interval and the first two

thirds of the next one are extracted as shown in Figure 6.4. To make all segments have the same length, each of them is uniformly resampled to a fixed length N, which in this case is fixed to 500. The segment with the highest correlation is taken as template for further analysis, from which the features proposed are extracted. However, many of these features require the exhaustive delineation of the characteristic loops.

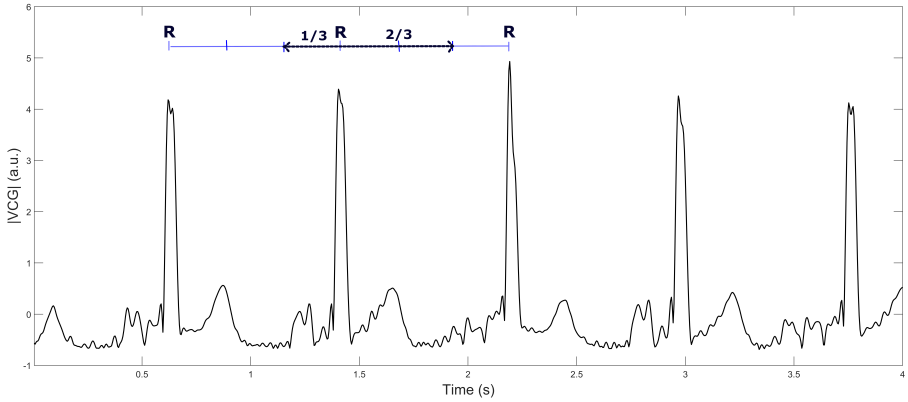

Figure 6.4: Example segmentation. Heartbeats in the VCG magnitude |VCG| are segmented taking 1/3 of the RR interval prior to the R-peak, and the first 2/3 of the next RR interval.

Accordingly, the R-peak is re-detected in this resampled segment using again the algorithm proposed in [122], and the Q and S waves are detected using the algorithm by Martínez et al. [116]. The T-peak and the T-end are identified using an adaptation of the algorithm by Zhang et al. [121, 206]. As indicated in Figure 6.3, these delineation is done in the VCG filtered up to 30Hz. However, the VCG loop used for further analysis should include higher frequencies to capture signs of fragmentation [64]. Therefore, the detected points are translated to the original VCG signal, bandpass-filtered between 0.6 and 70Hz.

6.3.2 Feature extraction

A combination of state-of-the-art and novel VCG features are proposed in this work in order to characterize the different patient populations under investigation. Except for the QRS-T angle, included due to its common use in VCG analysis, the rest of the features considered are only related to the QRS loop. This is done to enable a fair comparison with the evaluation of the LE framework proposed in Chapter 5, which

only included ventricle depolarization in the analysis. The features are listed in Table 6.1.

Table 6.1: Features derived from the VCG loops

Features		**Notation**
QRS-T angle		QRS-T angle
Percentage	Frontal	FQ1,FQ2,FQ3,FQ4
of QRS-loop area	Tranversal	TQ1,TQ2,TQ3,TQ4
per region	Sagittal	SQ1,SQ2,SQ3,SQ4
	Range	speedRange
	Standard Deviation	speedSD
Speed	Maximum Derivative	speedMaxDer
	Range Derivative	speedRangeDer
	SD Derivative	speedSDDer
	Range	curvRange
	Standard Deviation	curvSD
Curvature	Maximum Derivative	curvMaxDer
	Range Derivative	curvRangeDer
	SD Derivative	curvSDDer
	Frontal	FRF
Feret Ratio	Transversal	FRT
	Sagittal	FRS
QRS-axis		QRS-axis

Spatial QRS-T angle

One of the best-known VCG features is the angle between depolarization and repolarization vectors in the 3-D space, referred to as spatial QRS-T angle. This feature captures conduction abnormalities, associated with damaged myocardial regions causing ventricular rhythm disturbances [90].

The concept of QRS-T angle is shown in Figure 6.5. To calculate it, the Q, R and T points detected in the preprocessing stage are used. The first step is to define the vectors $\vec{qrs}$ and $\vec{t}$. The depolarization vector $\vec{qrs}$ connects the origin with the R-peak of the VCG, and it is shown in Figure 6.5 in blue. The origin corresponds to the start of the QRS loop [201]. The repolarization vector $\vec{t}$, in green, connects the origin with the T-peak. The spatial QRS-T angle is calculated using the normalized inner product of both vectors $\vec{qrs}$ and $\vec{t}$ in the 3-D space, as:

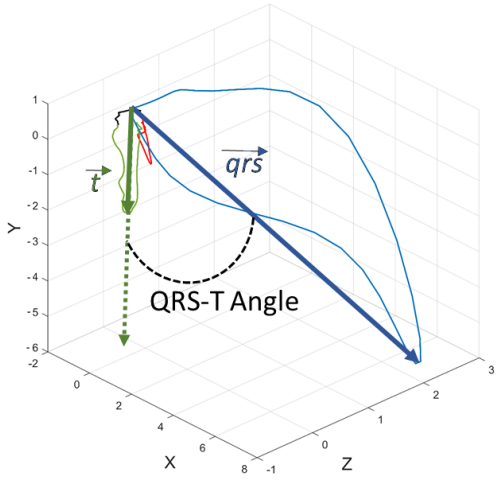

Figure 6.5: Graphic representation of the QRS-T angle. Figure based on [135]

$$QRS-T\,Angle = arccos\frac{\vec{qrs}\cdot\vec{t}}{|\vec{qrs}|\cdot|\vec{t}|}, \tag{6.2}$$

where $|\vec{qrs}|$ and $|\vec{t}|$ are the magnitudes of the $\vec{qrs}$ and $\vec{t}$ vectors.

Rijnbeek et al. determined the normal values of several ECG parameters for different age ranges and both for males and females in a study including more than 13000 12-lead ECG signals [145]. The normal QRS-T angle values for men and women between 60 and 69 years old correspond to 57º (2nd percentile 16, 98th percentile 125) and 52º (13 - 114) respectively. Wide QRS-T angles correspond to values $\geq$100 both for patients with IHD and DCM [61, 151].

Speed of the QRS loop

Arrhythmias are characterized by abnormal electrical conduction in the heart. These abnormalities can be identified by studying the speed of the cardiac vector along the QRS loop. The speed is defined as the distance covered by the heart vector to travel between two points, divided by the time taken to describe that trajectory. A study by Tereshchenko et al. proposed this analysis to differentiate between ECGs from healthy subjects and those of patients with sustained monomorphic ventricular tachycardia (smVT) and post-myocardial infarction (MI) [170]. Decreased conduction velocity along the ventricles caused by scar tissue is a known substrate of arrhythmogenesis [71].

While these changes are mostly perceptible invasively or using imaging techniques [8, 71], the speed metrics proposed by Tereshchenko have shown to capture this in a non-invasive way [170].

In line with their methodology, QRS loops were segmented in 10 segments of equal duration and the speed of the cardiac vector was calculated for each of these segments, as the distance covered by the cardiac vector in each segment divided by the duration of the segment. The distance covered at each of these segments is derived as:

$$dist = \sum_{i=2}^{N} d(x_i, x_{i-1}), \tag{6.3}$$

where N is the number of samples in each segment of the QRS loop and $d(x_i, x_{i-1})$ is the euclidean distance between two points in the 3D space. Here, also 5 metrics are derived from these 10 speed values, namely the range and the standard deviation; and the maximum, the range and the standard deviation of the derivative of the speed.

The normal speed trend along the QRS loop is defined by slow conduction at the beginning and end of the QRS complex, and the maximum velocity at the R-peak. An example of this pattern can be seen in Figure 6.6, where the QRS loop is shown. In blue, the VCG magnitude of the loop can be seen, and in red, the speed values. These values are calculated for 10 segments of equal length along the QRS loop.

The ventricular conduction in both IHD and DCM patients is expected to have a slower maximum in speed than that of normal ECGs, due to the presence of myocardial scar and the dilation of the ventricles, respectively, which interrupts normal cardiac conduction [170].

Curvature of the QRS loop

The morphology of the QRS loop can be characterized by its curvature, as proposed by Tereshchenko et al. [170]. This feature describes ventricular conduction, and in their study it was used to compare VCGs from healthy subjects, and patients with smVT or MI. The curvature c_i at each point x_i of a 3D vector was calculated as the inverse of the radius circumscribing points x_{i-1}, x_i and x_{i+1} [120]. Therefore, closed curves such as that one drawn by the R-peak in the QRS loop are circumscribed by small circles, i.e. small radius, that obtain large curvature values.

By obtaining the value of curvature for each point of the QRS loop, a detailed description of its fluctuations is provided. Nevertheless, such a local metric is very sensitive to errors in the QRS delineation, since noise and other artifacts around the Q and the S wave can lead to high values of curvature that do not contain relevant

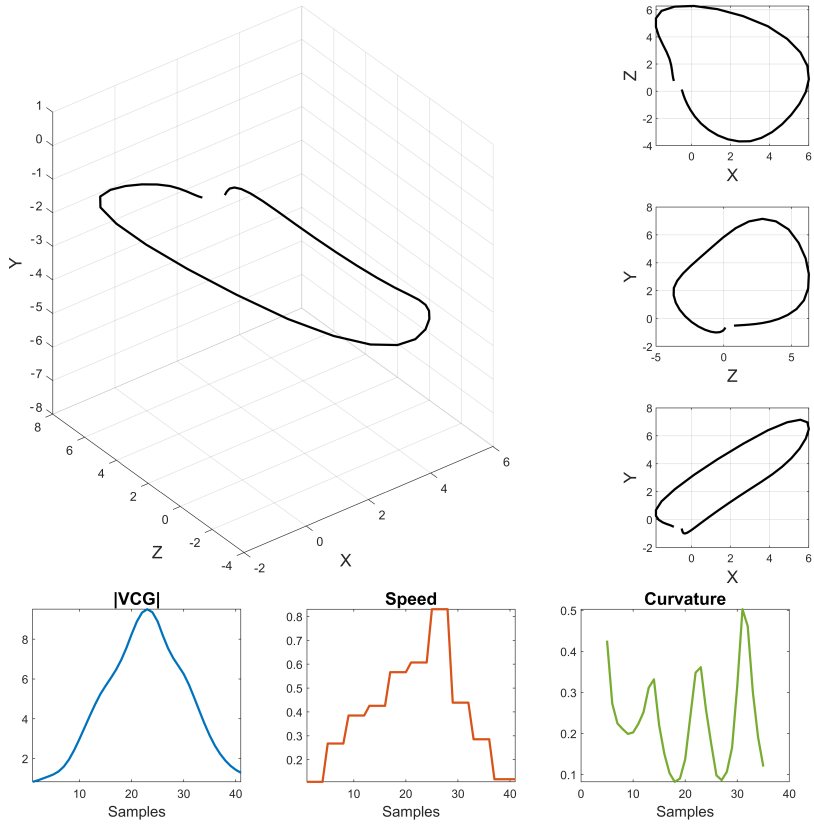

Figure 6.6: Example of a QRS loop with a normal pattern. The upper plots show the loop in the different anatomical planes. The lower plots show the magnitude of the VCG loop |VCG| (blue), the speed along the loop (red) and the curvature in the central 70% of the QRS loop (green).

information. Therefore, in order to remove potential segmentation inaccuracies, the 15% of the total length of the loop at its beginning and its end are removed from the curvature analysis. Hence, the values c_i were only obtained for the central 70% of the QRS loop. An example of the curvature can be seen in Figure 6.6, where the values of this feature along the QRS loop are shown in green.

Five metrics are proposed to characterize the variability of c_i along the loop, namely the range and the standard deviation of the curvature; and the maximum, the range, and the standard deviation of the derivative of the curvature. The results of Tereshschenko et al. indicate that healthy ECGs have a higher variability of curvature values, since the segments Q-R and R-S describe very open curves, while the R-peak presents the

maximum curvature. On the other hand, their VT patients presented less variability in the curvature values along the loop [170].

Feret ratio QRS loop

The curvature values associated to healthy VCGs indicate that the shape of the QRS loop for these VCGs is generally elongated, while QRS loops with delays in propagation are more constant in terms of curvature values [170]. Hence, this suggests that abnormalities in the ECG are translated into a rounded QRS loop in the VCG [60].

We propose the ratio between the maximum and minimum diameter of the QRS loop, which were calculated using the Feret's diameters of the loop for each of the anatomical planes, as a circularity metric. Feret's diameter (FD) is a technique used in digital image processing to analyze the shape and size of irregular particles [191]. It is defined as the distance between two parallel planes bounding the loop. This metric was calculated for each of the 3 anatomical planes. To calculate it, the QRS-loop in each of the planes was transformed in a black-and-white mask, as shown in Figure 6.7. After that, the maximum and minimum Feret diameters were calculated, and we define the Feret ratio (FR) as:

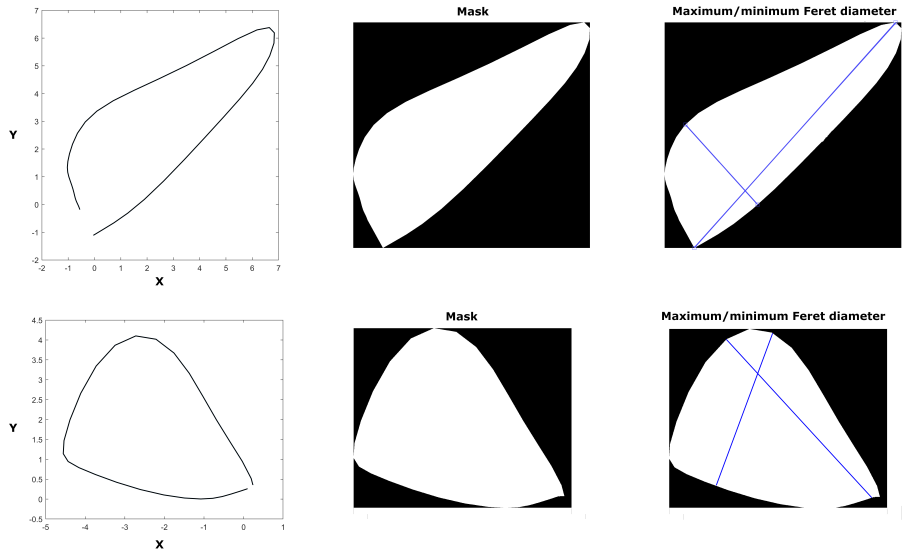

Figure 6.7: Steps to extract the Feret Ratio of two example loops.

$$FR = \frac{FD_{max}}{FD_{min}} \tag{6.4}$$

QRS area spatial distribution

The last set of features proposed are related to the direction of the QRS loop in the 3D space. By definition, the QRS loop is oriented towards the left, caudal, posterior direction of the body [145]. However, some studies have shown that ventricular conduction abnormalities could force the cardiac vector to alter its trajectory in the 3D space [60].

To characterize the trajectory of the loops in the 3D space, each of the anatomical planes was divided in 4 quadrants, as shown in Figure 6.8. For each of the planes, the area of the QRS loop contained in each of these quadrants was divided by the total area of the loop. The area per quadrant was defined as the surface limited by the axis defining the space and the contour in that plane. These calculations result in 12 features, 4 relative areas for each of the 3 anatomical planes.

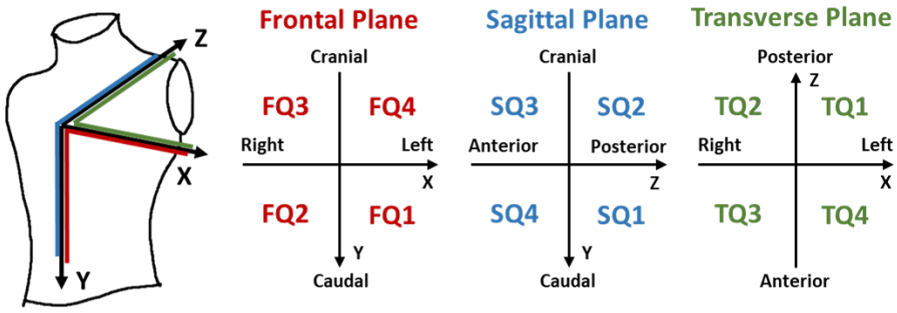

Figure 6.8: Scheme of the quadrants related to the anatomical planes.

QRS-axis

The previous features related to the percentage of area per quadrant present a novelty in VCG analysis, particularly for the sagittal and lateral plane. However, in the frontal plane this feature is similar to the well-described QRS-axis. This metric refers to the direction of the ventricular depolarization vector in the frontal plane. The QRS-axis is considered normal for adults when it is directed downwards and towards the left side of the body, creating an angle between -30º and 90º with lead I. Since the frontal plane can be described by the perpendicular leads I and aVF, the cardiac axis can be calculated as:

$$\text{QRS-axis} = arctg\left(\frac{2}{\sqrt{3}} \cdot \frac{\text{aVF}}{\text{I}}\right) \cdot \frac{180}{\pi}, \tag{6.5}$$

where aVF and I represent the net deflection of the QRS complex in those leads, derived as the addition of the amplitudes at the Q, R and S points [134]. Left axis deviation (LAD) is defined by a QRS-axis between -30º and -90º, and right axis deviation (RAD) 90º and 180º. While these ranges of values are orientative, deviation from the normal QRS-axis orientation can be caused by conduction disturbances, ventricular hypertrophy, and other underlying pathology [97, 156].

6.3.3 Experiments

The experiments performed in this analysis are similar to those described in Section 5.2.6:

- **Capturing characteristics of normal ECG**: The VCG features suggested were extracted for 10 groups of 32 subjects of the PTB-XL database selected at random. The goal of this experiment was to confirm the consistency of the features proposed, which should behave similarly for different normal ECGs.
- **Identifying cardiovascular diseases: differences between IHD and DCM**. The VCG features were evaluated in the characterization of structural cardiac abnormalities, by comparing them for three groups of subjects: normal ECGs, patients with IHD and patients with DCM.
- **Arrhythmia risk prediction for electrical storm**. These VCG features were applied for the analysis of arrhythmia prediction in ischemic and non-ischemic patients, for ES prediction. For this aim, the proposed VCG analysis was applied to normal subjects, and patients with and without history of ES.

6.3.4 Statistical Tests

The features of different groups of subjects were compared using the Kruskal-Wallis test with Bonferroni correction for multiple comparisons. Differences were considered significant for p-value < 0.05. Significant differences are illustrated in the violin plots using the $*$ sign. Additionally, the relation of these features with QRS duration were explored using Pearson correlation.

6.4 Results

6.4.1 Capturing characteristics of normal ECG

None of the VCG features was significantly different between the groups of normal ECGs. This suggests that they all belong to the same population, as it was also observed using the LE features (Section 5.3.1). However, the dispersion of the values for some of these VCG features was very large, much more than for LE. Figure 6.9 shows the violin plots for the QRS-T angle, for FQ1 and for the QRS axis. As it can be seen, all the groups cover a large range of values. Accordingly, experiments 2 and 3 were repeated 10 times in order to include the 10 groups of healthy subjects and confirm the consistency of the results.

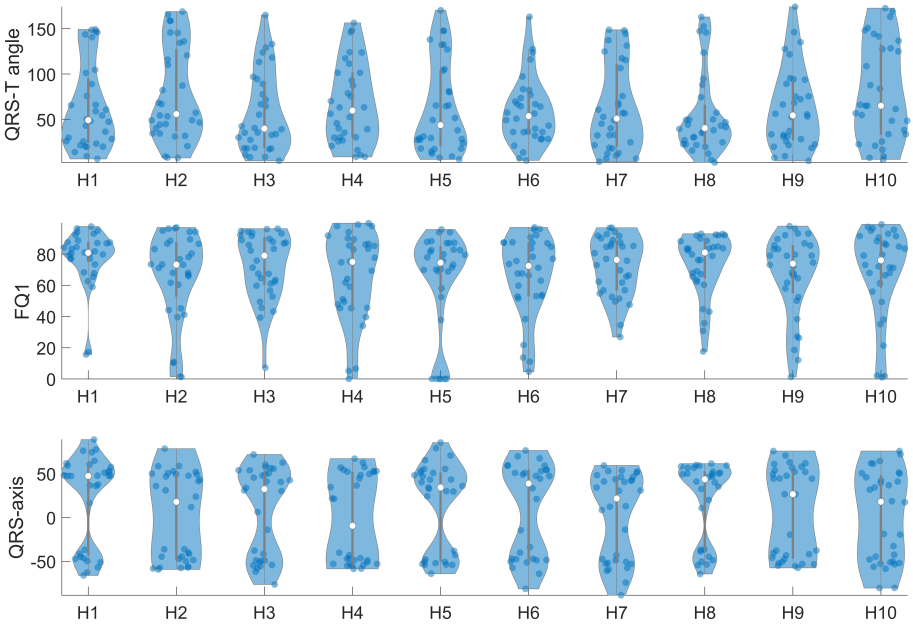

Figure 6.9: Violin plots of the normal ECGs for QRS-T angle, FQ1 and QRS-axis. The white dot represents the median of the values. No significant difference was found between the groups.

6.4.2 Identifying cardiovascular diseases: differences between IHD and DCM

The features obtained for each of the H_i, $i = 1,2,..10$ were compared to those of the IHD and DCM groups using Kruskal-Wallis test between the 3 populations. This test was repeated 10 times for each H_i to confirm the relevance of the findings.

The repetition of the tests for each H_i group provides a new dimension to the results, since it also evaluates the consistency of the metrics. Features that consistently identify the same results for each of the H_i groups can be considered more robust than those with contradictory results along the repetitions.

No feature was able to identify significant differences between IHD and DCM patients. In contrast, these differences were captured by the distance metrics and the curvature in LE loops, as explained in Section 5.3.2.

On the other hand, the only features that were able to discriminate between normal ECGs and both patient groups for all the iterations were those related to the speed in the QRS loop. These results show that normal VCG loops presented faster speeds than those of the patient groups. Due to their similar results, the correlation between the five speed features was calculated, obtaining values above 0.85 for all cases. Therefore, given their redundancy, only the standard deviation of the speed (speedSD) was considered for further analysis.

The speed was previously defined as the distance covered by the QRS vector divided by the time taken to cover it. Taking this into account, and given the significant differences identified between the QRS duration of the datasets under study, the relation between QRS duration and speed SD was explored for the 10 iterations. Pearson correlation coefficient ranged between -0.21 and -0.43, indicating a moderate relation between both characteristics. Figure 6.10 shows the feature space spanned by the speed SD and the QRS duration of the signals belonging to the group H_i corresponding to the largest Pearson correlation (ρ = -0.41). This figure confirms that wide QRS complexes are linked to slow speeds, and these are only present in ischemic and non-ischemic patients. Nevertheless, the speed values are higher for normal ECGs than for signals from patients with comparable QRS duration.

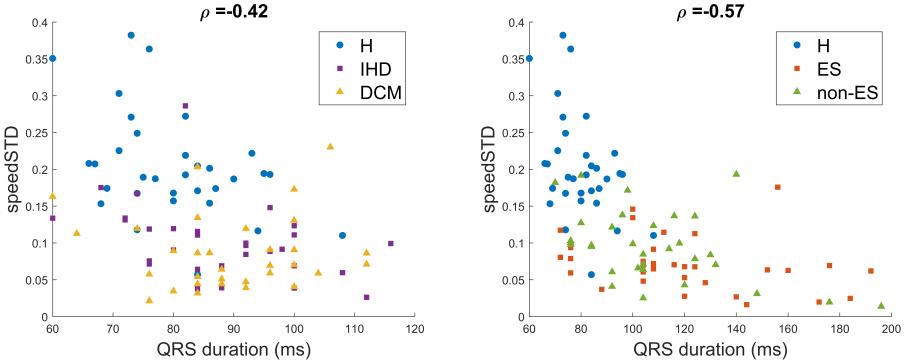

Figure 6.10: Correlation of QRS duration and speed SD. At the left, results for IHD and DCM patients (Section 6.4.2); at the right for ES and non-ES (Section 6.4.3).

The features that could identify significant differences between normal and IHD ECGs were the percentages of area in the IV quadrant for the horizontal and the transversal plane (HQ4 and TQ4), and the Feret ratio in the frontal plane. These results indicate that the QRS loops of normal subjects are mostly oriented towards the left side of the body, spread between quadrants 1 and 4; while the loops of IHD patients contain most of their area in the left posterior region, and more disperse in the transversal plane. Figure 6.11 shows the distribution of the areas for the different quadrants of the anatomical planes. The Feret ratio in the frontal plane is larger for normal subjects as shown in Figure 6.12, which suggests that IHD patients have more rounded loops in this plane. Differences between IHD and normal LE loops were also significant, as detected by the features associated to the first LE dimension and by the distance metrics.

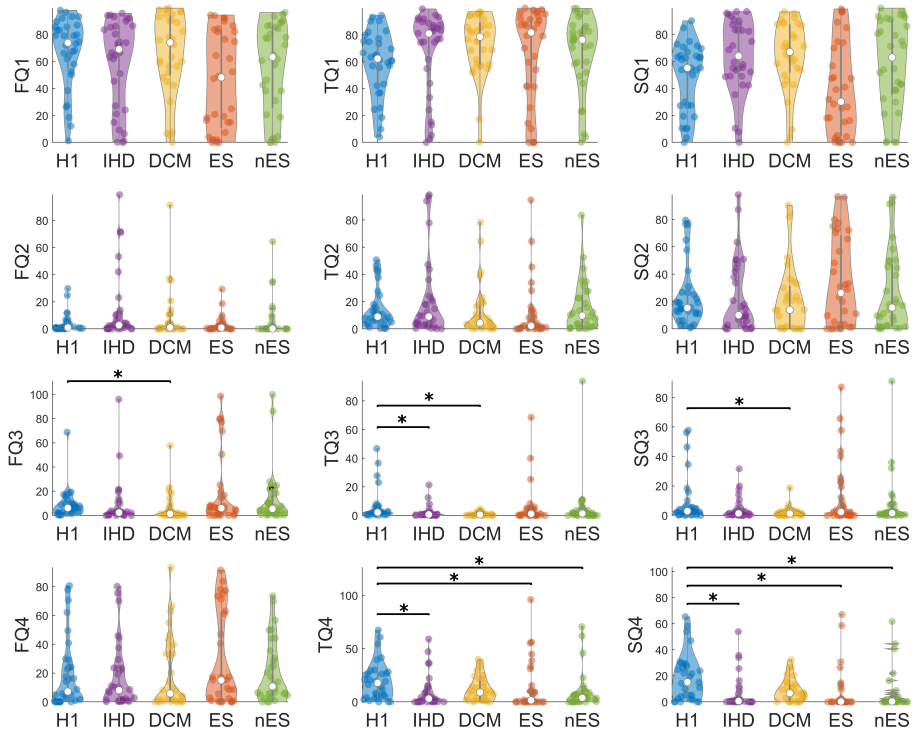

Figure 6.11: Distribution of the QRS-loops areas on the different quadrants.

Regarding differences between normal and DCM loops, the features that more consistently identified differences between these two groups were the spatial QRS-T angle and the percentage of area in FQ3, which were found significantly different in 9 out of the 10 iterations for the different groups of normal VCGs. The QRS-T angle was wider for DCM patients than for normal subjects, as it can be seen in Figure 6.12, and the percentage of area in FQ3 for normal ECGs was lower than for DCM patients. These differences were also identified using the LE approach, with the features associated to the first dimension of the LE space and the curvature.

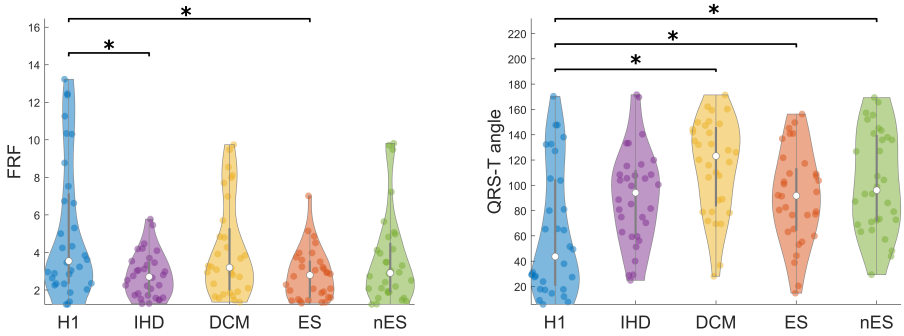

Figure 6.12: Violin plots for the Feret ratio in the frontal plane (FRF) and the QRS-T angle.

In contrast to the previous results, the five curvature features and part of the features describing the distribution of the area in the quadrants, namely FQ1, FQ2, FQ4, HQ2, TQ2 and TQ3, were not different between normal and cardiomyopathy patients for any of the H_i groups. Features like the QRS-axis or the Feret ratio in the sagittal and lateral plane obtained inconsistent results for the 10 repetitions with different H_i groups. This indicates that the differences between the studied populations are not significant for these features, and it confirms their wide variability depending on the subjects under analysis.

6.4.3 Arrhythmia prediction: Electrical storm

None of the VCG features could distinguish between ES and non-ES patients, indicating that there are no clear differences between these two groups in the visit corresponding to 4 months before the storm event.

The features that did not show significant differences in any of the 10 iterations were the Feret ratio in the transversal and sagittal plane (FRT, FRS) and the distribution of the QRS-loops area in 1, 2 and 3 quadrants of the three anatomical plane. Additionally, the QRS-axis only identified differences between two of the H_i groups and ES ECGs.

Similarly to the cases with structural abnormalities, the QRS-loops of non-ES patients did not show consistent significant differences in terms of curvature compared to those of healthy subjects. However, these features identified significant differences between some of the H_i groups and the ES patients. For the SD of the curvature, and the SD of the derivative of the curvature, these differences were identified in 6 out of 10 iterations. However, these results are not conclusive due to their lack of consistency along the 10

repetitions. The variability of these results may be due to the sensitivity to outliers of the curvature, as it can be seen in Figure 6.13.

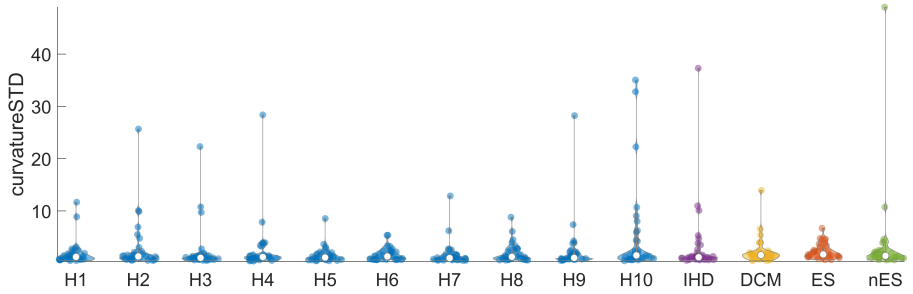

Figure 6.13: Curvature values for the 10 H_i groups and for the patients.

In this case, the features that always identified differences between the patients and normal signals were all the speed features and the percentages of QRS-loop laying in quadrant 4 of the transversal and sagittal plane (TQ4, SQ4). Similarly to the previous experiment, the 5 speed features were correlated between them, obtaining values above 0.85. Using speedSD as reference of the speed features, its correlation with the QRS duration of the signals was calculated, obtaining a ρ = -0.57. These values, higher than in the previous experiment, are related to the presence of broader QRS signals in the ES and non-ES patients than in the rest of the signals. Nevertheless, the values of speedSD for signals of comparable QRS duration (below 100 ms) are lower for ES and non-ES patients than for the healthy group.

The features related to the distribution of the loops in the 4th quadrant of the transversal and sagittal plane present similar results than those of IHD patients. Normal QRS-loops distribute part of their surface in these quadrants, while ES and non-ES groups barely do.

For this experiment, the results are similar to those identified with the LE features, which detected significant differences between normal ECGs and those of patients, but not between patients with and without history of ES.

6.5 Discussion

In this relatively small sample none of the tested VCG features could identify differences between the ECGs of IHD and DCM patients, and neither between those manifesting with ES or not. This in contrast to the data-driven LE approach that did find significant differences between IHD and DCM, but also not for ES.

However, similarly to LE features, the VCG analysis found differences between normal and cardiomyopathy patients. In particular, the features related to the speed of the heart vector along the QRS-loop were able to find significant differences between healthy subjects and each of the patient groups, for all the normal groups H_i. According to the correlation analysis between the speed features, all of them were correlated, and therefore they all captured the same information. By analyzing only the standard deviation of the speed along the QRS loop it was shown that this feature is moderately correlated to the QRS duration. However, for comparable QRS durations, it still was obvious that normal VCG loops presented faster speeds than those of the different groups of patients.

These speed features and those related to the curvature of the QRS-loop were inspired by the work by Tereshchenko et al. [170]. However, the curvature features could not distinguish between VCGs of different groups of patients. These metrics are much more sensitive to outliers than those related to speed, since the curvature is calculated for each point in the trajectory, while the speed is calculated per sections. These features were meant to be that sensitive in the first place, since they were expected to capture loop foldings and other twists that could be related to fragmentation in the QRS [155]. However, while these features captured relevant information when applied to LE, for VCG these variations do not seem to contain information. This could be explained by the way the 12-lead information is integrated in each of the dimensionality reduction approaches. In VCG, the uniform way of reducing the dimensionality of all signals, without considering their specifications, can lead to potential information loss or the inclusion of irrelevant information to the low-dimensional data, such as noise [112]. In fact, some authors have suggested patient-based or disease-based transformations for VCG [40, 189]. On the other hand, the data-driven dimensionality reduction performed by the LE algorithm can smartly combine the information in the multi-lead signal, responding to the characteristics of the signal.

Regarding the shape of the loops in the different planes, only the FRF could detect differences between normal and IHD signals, and between normal and ES. This suggests that only the roundness in the frontal plane can capture conduction abnormalities in the VCG. Current clinical analysis mostly explores the behavior of the heart vector in this plane, which is the one presenting the largest trajectories [60, 151, 138]. The study by Ghista et al. does report similar results, indicating that rounder loops are associated to myocardial infarction and conduction abnormalities [60].

The work by Ghista et al. was also the inspiration to explore the distribution of the loops in the different quadrants. While the study in [60] only explored the frontal plane, here it was shown that the conduction pathways significantly differ also for the sagittal and transversal plane. The results show that for IHD, ES and non-ES patients, their loops barely lie on the 4th quadrant of the transversal plane, while healthy and DCM loops cover more of these regions.

While the behavior of the area distribution was slightly related to the QRS-axis according to the analysis performed in [24], in the present analysis these differences were not identified. This can be due to the presence of different QRS-axis orientations in the PTB-XL database, according to the annotations [190].

Given the relevance of the QRS-T angle in VCG analysis, and especially in arrhythmia prediction, this feature was also calculated for the different groups. While in general the QRS-T angle was larger for all the patient groups than for healthy subjects, this feature took a wide range of values for the latter, as shown in Figure 6.9. This may be due to the calculation of the angle using the maximal QRS and T vectors, which is more sensitive to outliers than other formulations.

As discussed in Section 6.2, several studies have explored the morphology, orientation, dispersion and beat-to-beat variability of VCG loops for different pathologies and outcome predictions. In this study, two key fields of VCG analysis, namely T-loop descriptors and beat-to-beat variability have not been addressed. Future work should consider the inclusion of the T-wave, both in LE and VCG analysis, especially for the identification of arrhythmia risk biomarkers [103, 202]. Additionally, longer recordings would allow an analysis of beat-to-beat variability, which according to the literature would contain more information about irregular conduction of the heart.

Nevertheless, the analysis performed in this and the previous chapter allow identifying the strengths of VCG and LE in a comparable set-up including exclusively the QRS complex. The study of ventricular depolarization allows to reduce the impact of delineation errors, which are more common in the analysis of the T-wave [201], and unravel the describing abilities of LE and VCG. While one of the advantages of VCG over ECG was the possibility of extracting features less sensitive to signal delineation, like the QRS-T angle, the Feret ratio, or the speed; this was even more remarkable for LE. Future studies should compare both approaches for different signal contamination levels.

One of the methods to derive an orthogonal 3D representation from 12-lead ECG applies SVD and keeps the three first singular vectors to describe loops [1]. While this approach shares some similarities with the LE algorithm, it does not exploit the relationship between consecutive samples since it does not include the calculation of the Laplacian of the similarity matrix. Nevertheless, this SVD-based SVD-based 3D representation offers a data-driven alternative to classic transformation-based VCG approaches which should be compared to our LE approach in future work.

6.6 Conclusion

In this chapter, a selection of VCG-based features has been applied for the analysis of normal ECGs, the study of structural abnormalities by comparing IHD and DCM VCGs and the analysis of arrhythmogenicity-related changes for the analysis of ES.

While no significant differences were identified between random groups of normal VCGs, indicating a robust feature extraction consistent for different healthy subjects; no differences were identified neither between the different patient groups. While the underlying pathology and propagation patterns of IHD and DCM patients should differ according to the literature and to the findings of the LE features proposed in Chapter 6, these differences were not captured by the VCG analysis. This was neither the case for the ES and non-ES patients, but this remained also an open question from the previous chapter.

Nevertheless, the speed of the QRS-loop, and the percentage of QRS-loop area covering the 4th quadrant in the transversal and sagittal plane, identified differences between normal VCGs and those from the different patient groups.

The results presented in this and the previous chapter suggest that the LE approach might be more robust to outliers than VCG. They confirm the advantages of data-driven approaches in dimensionality reduction of 12-lead ECG information, which reduces the influence of artefacts and errors in signal delineation, setting the path for further integrated analysis of multi-dimensional signals.

Chapter 7

Manifold learning for unsupervised feature selection: the U2FS algorithm

Data-driven models are becoming a very powerful tool in the transformation of healthcare, since they allow extracting relevant information to support decision-making. These studies often involve the use of high-dimensional data, which analysis is extremely challenging since often not all the included characteristics are relevant for the problem under investigation. Feature selection techniques are very useful approaches for dimensionality reduction in data analysis. They provide interpretable results by reducing the dimensions of the data to a subset of the original set of features. When the data lack annotations, unsupervised feature selectors are required for their analysis. Several algorithms have been proposed for this purpose, but despite their large applicability, they can be very inaccessible or cumbersome to use, mainly due to the need for tuning non-intuitive parameters and to their high computational demands.

In this chapter, a publicly available ready-to-use unsupervised feature selector is proposed, with comparable results to the state-of-the-art at a much lower computational cost. The algorithm, based on the utility metric for unsupervised feature selection (U2FS) was evaluated in simulated data and benchmark databases, and its performance was comparable to the state-of-the-art but with a much lower computation cost. This work was published in:

> **Villa, A.**, Mundanad Narayanan A, Van Huffel S, Bertrand A, Varon C. 2021. Utility metric for unsupervised feature selection. *PeerJ Computer Science* 7:e477

7.1 Introduction

Many applications of data science require the study of highly multi-dimensional data. Some examples are the study of multi-lead ECG, as mentioned in the previous chapters, or other clinical applications where medication, demographics or other features extracted from biomedical signals or clinical records may be used for analysis. In these applications, one patient is represented by a long set of variables, where each variable represents one dimension.

A high number of dimensions implies a high computational cost as well as a large amount of memory required. Furthermore, this often leads to problems related to the curse of dimensionality [179] and thus, to irrelevant and redundant data for machine learning algorithms [5]. Therefore, it is crucial to perform dimensionality reduction before analyzing the data.

There are two types of dimensionality reduction techniques. So-called feature selection techniques directly select a subset of the original features. On the other hand, transformation techniques compute a new (smaller) set of features, each of which are derived from all features of the original set. Some examples of these are Principal Component Analysis (PCA) [192], Independent Component Analysis (ICA) [88] or the Extended Sammon Projection (ESP) [3]. While these methods lead to a reduction in the number of dimensions, results are less interpretable, since their direct relationship with the original set of features is lost.

In this work, the focus is on unsupervised feature selectors. Since these methods do not rely on the availability of labels or annotations in the data, the information comes from the learning of the underlying structure of the data. Despite this challenge, the generalization capabilities of these methods are typically better than for supervised or semi-supervised methods [69]. Within unsupervised feature selectors, sparse learning based methods have gained attention in the last 20 years [100]. These methods rely on graph theory and manifold learning to learn the underlying structures of the data [107], and they apply sparsity inducing techniques to perform subset selection. However, to the best of our knowledge, none explores specifically the behavior of these methods with data presenting non-linear relationships between the features (i.e., dimensions). While the graph definition step can make use of kernels to tackle non-linearities, these can be heavily affected by the curse of dimensionality, since they are often based on a distance metric [2].

After the manifold learning stage, sparse regression is applied to score the relevance of the features in the structures present in the graph. These formulations make use of sparsity-inducing regularization techniques to provide the final subset of features selected, and thus, they are highly computationally expensive. These methods are often referred to as structured sparsity-inducing feature selectors (SSFS), or sparse learning

based methods [67][100].

Despite the large amount of unsupervised SSFS algorithms described in the literature, these methods are cumbersome to use for a novice user. This is not only due to the codes not being publicly available, but also due to the algorithms requiring regularization parameters which are difficult to tune, in particular in unsupervised settings.

In this work, an efficient unsupervised feature selector based on the utility metric (U2FS) is proposed. U2FS is a ready-to-use, publicly available unsupervised sparsity-inducing feature selector designed to be robust for data containing non-linearities. The code is available here: https://github.com/avillago/u2fs, where all functions and example codes are published. The main contributions of this work are:

- The definition of a new method to automatically approximate the radial-basis function (RBF) kernel parameter without the need for a user-defined tuning parameter. This method is used to tackle the curse of dimensionality when embedding the data taking non-linearities into account.
- The suggestion of a backwards greedy approach for the stage of subset selection, based on the utility metric for the least-squares problem. The utility metric was proposed in the framework of supervised learning [18], and has been used for channel selection in applications such as electroencephalography (EEG) [130], sensor networks [166], and microphone arrays [165]. Nevertheless, this is the first work in which this type of approach is proposed for the sparsity-inducing stage of feature selection.
- Propose a non-parametric and efficient unsupervised SSFS algorithm. This work analyzes the proposed method U2FS in terms of its complexity, and of its performance on simulated and benchmark data. The goal is to reduce the computational cost while maintaining a comparable performance with respect to the state-of-the-art. In order to prove this, U2FS is compared to three related state-of-the-art algorithms in terms of accuracy of the features selected, and computational complexity of the algorithm.

The rest of the chapter is structured as follows. In Section 7.2, previous algorithms on SSFS are summarized. In Section 7.3, the proposed U2FS method is described: first the manifold learning stage, together with the algorithm proposed for the selection of the kernel parameter; and further on, the utility metric is discussed and adapted to feature selection. The experiments performed in simulations and benchmark databases, as well as the results obtained are described in Sections 7.4 and 7.5. Finally, the last section provides some conclusions.

7.2 Related Work

Sparsity-inducing feature selection methods have become widely used in unsupervised learning applications for high-dimensional data. This is due to two reasons. On the one hand, the use of manifold learning guarantees the preservation of local structures present in the high-dimensional data. Additionally, its combination with feature selection techniques not only reduces the dimensionality of the data, but also guarantees interpretability.

Sparsity-inducing feature selectors learn the structures present in the data via connectivity graphs obtained in the high-dimensional space [195]. The combination of manifold learning and regularization techniques to impose sparsity, allows the selection of a subset of features from the original dataset that are able to describe these structures in a smaller dimensional space.

These algorithms make use of sparsity-inducing regularization approaches to stress those features that are more relevant for data separation. The sparsity of these approaches is controlled by different statistical norms ($l_{r,p}$-norms), which contribute to the generalization capability of the methods, adapting them to binary or multi-class problems [67]. One drawback of these sparse regression techniques is that generally, they rely on optimization methods, which are computationally expensive.

The Laplacian Score [75] was the first method to perform spectral feature selection in an unsupervised way. Based on the Laplacian obtained from the spectral embedding of the data, it obtains a score based on locality preservation. SPEC [210] is a framework that contains this previous approach, but it additionally allows for both supervised or unsupervised learning, including other similarity metrics, as well as other ranking functions. These approaches evaluate each feature independently, without considering feature interactions. These interactions are, however, taken into account in Multi-Cluster Feature Selection (MCFS) [26], where a multi-cluster approach is defined based on the eigendecomposition of a similarity matrix. The subset selection is performed applying an l_1-norm regularizer to approximate the eigenvectors obtained from the spectral embedding of the data inducing sparsity. In UDFS [196] the l_1-norm regularizer is substituted by a $l_{2,1}$-norm to apply sample and feature-wise constraints, and a discriminative analysis is added in the graph description. In NDFS [101], the use of the $l_{2,1}$-norm is preserved, but a non-negative constraint is added to the spectral clustering stage. Additionally, this algorithm performs feature selection and spectral clustering simultaneously.

The aforementioned algorithms perform manifold learning and subset selection in a sequential way. However, other methods tackle these simultaneously, in order to adaptively change the similarity metric or the selection criteria regarding the error obtained between the original data and the new representation. Examples of these

algorithms are JELSR [80], SOGFS [133], (R)JGSC [211] and DSRMR [168], and all make use of an $l_{2,1}$-norm. Most recently, the SAMM-FS algorithm was proposed [207], where a combination of similarity measures is used to build the similarity graph, and the $l_{2,0}$-norm is used for regression. This group of algorithms are currently the ones achieving the best results, at the cost of using complex optimization techniques to adaptively tune both stages of the feature selection process. While this can lead to good results, it comes with a high computation cost, which might hamper the tuning process, or might simply not be worthy for some applications. SAMM-FS and SOGFS are the ones that more specifically suggest new approaches to perform the embedding stage, by optimally creating the graph [133] or deriving it from a combination of different similarity metrics [207]. Again, both approaches require computationally expensive optimization techniques to select a subset of features.

In summary, even if SSFS methods are getting more sophisticated and accurate, this results in algorithms becoming more complex in terms of computation time, and in the ease of use. The use of advanced numerical optimization techniques to improve results makes algorithms more complex, and requires regularization parameters which are not easy to tune. In this work, the combination of a new approach to estimate the graph connectivity based on the RBF kernel, together with the use of the utility metric for subset selection, results in an efficient SSFS algorithm, which is easy to use and with lower complexity than the state-of-the-art. This efficient implementation is competitive with state-of-the-art methods in terms of performance, while using a simpler strategy, which is faster to compute and easier to use.

7.3 Methods

This section describes the proposed U2FS algorithm, which focuses on selecting the relevant features in an unsupervised way, at a relatively small computational cost. The method is divided in three parts. Firstly, the suggested manifold learning approach is explained, where an embedding based on binary weighting and the RBF kernel are used. Then a method to select the kernel parameter of the RBF kernel is proposed, specially designed for high-dimensional data. Once the manifold learning stage is explained, the Utility metric is proposed as a new approach for subset selection.

7.3.1 Manifold learning considering non-linearities

Given a data matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, with $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_N]$, $\mathbf{x}_i = [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(d)}]$, $i = 1, \ldots, N$, N the number of data points, and d the number of features (i.e., dimensions) in the data. The aim is to learn the structure hidden in the d-dimensional data and approximate it with only a subset of the original features. In this work, this structure will

be identified by means of clustering, where the dataset is assumed to be characterized by c clusters.

In spectral clustering, the clustering structure of this data can be obtained by studying the eigenvectors derived from a Laplacian built from the original data ([186], [19]). The data is represented using a graph $G = (\mathscr{V}, \mathscr{E})$. $\mathscr{V}$ is the set of vertices $\mathbf{v}_i$, $i = 1, \ldots, N$ where $\mathbf{v}_i = \mathbf{x}_i$. $\mathscr{E} = \{e_{ij}\}$ with $i = 1, \ldots, N$ $j = 1, \ldots, N$ is the set of edges between the vertices where $\{e_{ij}\}$ denotes the edge between vertices v_i and v_j. The weight of these edges is determined by the entries $w_{ij} \geq 0$ of a similarity matrix $\mathbf{W}$. We define the graph as undirected. Therefore, the similarity matrix $\mathbf{W}$, is symmetric (since $w_{ij} = w_{ji}$, with the diagonal set to $w_{ii} = 0$).

Typically, $\mathbf{W}$ is computed after coding the pairwise distances between all N data points. There are several ways of doing this, such as calculating the k-nearest neighbours (KNN) for each point, or choosing the ε-neighbors below a certain distance [15].

In this work, two similarity matrices are adopted inspired by the work in [26], namely a binary one and one based on an RBF kernel. The binary weighting is based on KNN, being $w_{ij} = 1$ if and only if vertex i is within the K closest points to vertex j. Being a non-parametric approach, the binary embedding allows to simply characterize the connectivity of the data.

Additionally, the use of the RBF kernel is considered, which is well suited for non-linearities and allows the characterization of complex and sparse structures [186]. The RBF kernel is defined as $K(\mathbf{x}_i, \mathbf{x}_j) = exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma^2)$. The selection of the kernel parameter σ is a long-standing challenge in machine learning. For instance, in [26], σ^2 is defined as the mean of all the distances between the data points. Alternatively, a rule of thumb, uses the sum of the standard deviations of the data along each dimension [178]. However, the estimation of this parameter is highly influenced by the amount of features or dimensions in the data, making it less robust to noise and irrelevant features. In the next section, a new and better informed method to approximate the kernel parameter is proposed.

The graph G, defined by the similarity matrix $\mathbf{W}$, can be partitioned into multiple disjoint sets. Given the focus on multi-cluster data of our approach, the k-Way Normalized Cut (*NCut*) Relaxation is used, as proposed in [132]. In order to obtain this partition, the degree matrix $\mathbf{D}$ of $\mathbf{W}$ must be calculated. $\mathbf{D}$ is a diagonal matrix for which each element on the diagonal is calculated as $D_{ii} = \sum_j W_{i,j}$. The normalized Laplacian $\mathbf{L}$ is then obtained as $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, as suggested in [186]. The vectors $\mathbf{y}$ embedding the data in $\mathbf{L}$ can be extracted from the eigenvalue problem [30]:

$$\mathbf{L}\mathbf{y} = \lambda\mathbf{y} \tag{7.1}$$

Given the use of a normalized Laplacian for the data embedding, the vectors $\mathbf{y}$ must be adjusted using the degree matrix $\mathbf{D}$:

$$\alpha = \mathbf{D}^{1/2}\mathbf{y}, \tag{7.2}$$

which means that α is the solution of the generalized eigenvalue problem of the pair $\mathbf{W}$ and $\mathbf{D}$. These eigenvectors α are a new representation of the data, that gathers the most relevant information about the structures appearing in the high-dimensional space. The c eigenvectors, corresponding to the c highest eigenvalues (after excluding the largest one), can be used to characterize the data in a lower dimensional space [132]. Thus, the matrix $\mathbf{E} = [\alpha_1, \alpha_2, \ldots, \alpha_c]$ containing column-wise the c selected eigenvectors, will be the low-dimensional representation of the data to be mimicked using a subset of the original features, as suggested in [26].

7.3.2 Kernel parameter approximation for high-dimensional data

One of the most used similarity functions is the RBF kernel, which allows exploring non-linearities in the data. Nevertheless, the kernel parameter σ^2 must be selected correctly, to avoid overfitting or the allocation of all data points to the same cluster. This work proposes a new approach to approximate this kernel parameter, which will be denoted by $\hat{\sigma}^2$ when derived from our method. This method takes into account the curse of dimensionality and the potential irrelevant features or dimensions in the data.

As a rule of thumb, σ^2 is approximated as the sum of the standard deviation of the data along each dimension [178]. This approximation grows with the number of features (i.e. dimensions) of the data, and thus, it is not able to capture its underlying structures in high-dimensional spaces. Nevertheless, this σ^2 is commonly used as an initialization value, around which a search is performed, considering some objective function [4, 178].

The MCFS algorithm skips the search around an initialization of the σ^2 value by substituting the sum of the standard deviations by the mean of these [26]. By doing so, the value of σ^2 does not overly grow. This estimation of σ^2 suggested in [26] will be referred to as σ_0^2. A drawback of this approximation in high-dimensional spaces is that it treats all dimensions as equally relevant for the final estimation of σ_0^2, regardless of the amount of information that they actually contain.

The aim of the proposed approach is to provide a functional value of σ^2 that does not require any additional search, while being robust to high-dimensional data. Therefore, this work proposes an approximation technique based on two factors: the distances between the points, and the number of features or dimensions in the data.

The most commonly used distance metric is the euclidean distance. However, it is very sensitive to high-dimensional data, deriving unsubstantial distances when a high number of features is involved in the calculation [2]. In this work, the use of the Manhattan or taxicab distance [144] is proposed, given its robustness when applied to high-dimensional data [2]. For each feature l, the Manhattan distance δ_l is calculated as:

$$\delta_l = \frac{1}{N} \sum_{i,j=1}^{N} |x_{il} - x_{jl}| \tag{7.3}$$

Additionally, in order to reduce the impact of irrelevant or redundant features, a system of weights is added to the approximation of $\hat{\sigma}^2$. The goal is to only take into account the distances associated to features that contain relevant information about the structure of the data. To calculate these weights, the probability density function (PDF) of each feature is compared with a Gaussian distribution. Higher weights are assigned to the features with less Gaussian behavior, i.e. those the PDF of which differs the most from a Gaussian distribution. By doing so, these will influence more the final $\hat{\sigma}^2$ value, since they allow a better separation of the structures present in the data.

Figure 7.1 shows a graphical representation of this estimation. The dataset in the example has 3 dimensions or features: f_1, f_2 and f_3. f_1 and f_2 contain the main clustering information, as it can be observed in Figure 7.1a, while f_3 is a noisy version of f_1, derived as $f_3 = f_1 + 1.5n$, where n is drawn from a normal distribution $\mathcal{N}(0,1)$. Figures 7.1b, 7.1c and 7.1d show in a continuous black line the PDFs derived from the data, and in a grey dash line their fitted Gaussian, in dimensions f_1, f_2 and f_3 respectively. This fitted Gaussian was derived using the Curve Fitting toolbox of Matlab™. As it can be observed, the matching of a Gaussian with an irrelevant feature is almost perfect, while those features that contain more information, like f_1 and f_2, deviate much more from a normal distribution.

Making use of these differences, an error, denoted ϕ_l, for each feature l, where $l = 1,\ldots,d$, is calculated as:

$$\phi_l = \frac{1}{H} \sum_{i=1}^{H} (p_i - g_i)^2, \tag{7.4}$$

where H is the number of bins in which the range of the data is divided to estimate the PDF (p), and g is the fitted Gaussian. The number of bins in this work is set to 100 for standardization purposes. Equation (7.4) corresponds to the mean-squared error (MSE) between the PDF of the data over feature l and its fitted Gaussian. From these ϕ_l, the final weights b_l are calculated as:

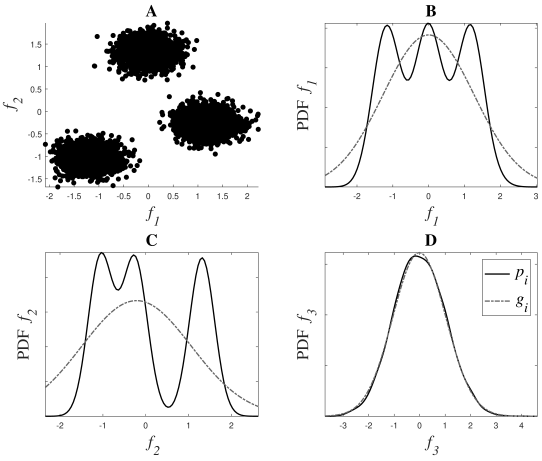

Figure 7.1: Weight system for relevance estimation. In Figure 7.1A, f_1 and f_2 can be seen. 7.1B, 7.1C and 7.1D show in black the PDFs p_i of f_1, f_2 and f_3 respectively, and in grey dotted line their fitted Gaussian g_i.

$$b_l = \frac{\phi_l}{\sum_{l=1}^{d} \phi_l} \tag{7.5}$$

Therefore, combining (7.3) and (7.5), the proposed approximation, denoted $\hat{\sigma}^2$, is derived as:

$$\hat{\sigma}^2 = \sum_{l=1}^{d} b_l \delta_l, \tag{7.6}$$

which gathers the distances present in the most relevant features, giving less importance to the dimensions that do not contribute to describe the structure of the data. The complete algorithm to calculate $\hat{\sigma}^2$ is described in Algorithm 3.

Algorithm 3 Kernel parameter approximation for high-dimensional data.

Input: Data $\mathbf{X} \in \mathbb{R}^{N \times d}$.
Output: Sigma parameter $\hat{\sigma}^2$

1: Calculate the Manhattan distances between the datapoints using Equation (7.3): vector of distances per feature δ_l.
2: Obtain the weights for each of the features using Equations (7.4) and (7.5): weights b_l.
3: Calculate $\hat{\sigma}^2$ using Equation (7.6).

7.3.3 Utility metric for feature subset selection

In the manifold learning stage, a new representation $\mathbf{E}$ of the data based on the eigenvectors was built, which described the main structures present in the original high-dimensional data. The goal is to select a subset of the features which best approximates the data in this new representation. In the literature, this feature selection problem is formulated using a graph-based loss function and a sparse regularizer of the coefficients is used to select a subset of features, as explained in [211]. The main idea of these approaches is to regress the data to its low dimensional embedding along with some sparse regularization. The use of such regularization techniques reduces overfitting and achieves dimensionality reduction. This regression is generally formulated as a least squares (LS) problem, and in many of these cases, the metric that is used for feature selection is the magnitude of their corresponding weights in the least squares solution [26, 67]. However, the optimized weights do not necessarily reflect the importance of the corresponding feature as it is scaling dependent and it does not properly take interactions across features into account [18]. Instead, the importance of a feature can be quantified using the increase in least-squared error (LSE) if that feature was to be removed and the weights were re-optimized. This increase in LSE, called the 'utility' of the feature can be efficiently computed [18] and can be used as an informative metric for a greedy backwards feature selection procedure [18, 130, 166], as an alternative for (group-)LASSO based techniques. Under some technical conditions, a greedy selection based on this utility metric can even be shown to lead to the optimal subset [35].

After representing the dataset using the matrix $\mathbf{E} \in \mathbb{R}^{N \times c}$ containing the c eigenvectors, the following LS optimization problem finds the weights $\mathbf{p}$ that best approximate the data $\mathbf{X}$ in the c-dimensional representation in $\mathbf{E}$:

$$J = \min_{\mathbf{P}} \frac{1}{N} ||\mathbf{X}\mathbf{p} - \mathbf{E}||_F^2 \tag{7.7}$$

where J is the cost or the LSE and $||.||_F$ denotes the Frobenius norm.

If $\mathbf{X}$ is a full rank matrix and if $N > d$, the LS solution $\hat{\mathbf{p}}$ of (7.7) is

$$\hat{\mathbf{p}} = \mathbf{R}_{\mathbf{XX}}^{-1}\mathbf{R}_{\mathbf{XE}}, \tag{7.8}$$

with $\mathbf{R}_{\mathbf{XX}} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$ and $\mathbf{R}_{\mathbf{XE}} = \frac{1}{N}\mathbf{X}^T\mathbf{E}$.

The goal of this feature selection method is to select the subset of $s(< d)$ features that best represents $\mathbf{E}$. This feature selection problem can be reduced to the selection of the best $s(< d)$ columns of $\mathbf{X}$ which minimize (7.7). However, this is inherently a combinatorial problem and is computationally unfeasible to solve. Nevertheless, several greedy and approximative methods have been proposed [67, 133, 130]. In the current work, the use of the utility metric for subset selection is proposed to select these best s columns.

The utility of a feature l of $\mathbf{X}$, in an LS problem like (7.7), is defined as the increase in the LSE J when the column corresponding to the l-th feature in $\mathbf{X}$ is removed from the problem and the new optimal weight matrix, $\hat{\mathbf{p}}_{-l}$, is re-computed similar to (7.8). Consider the new LSE after the removal of feature l and the re-computation of the weight matrix $\hat{\mathbf{p}}_{-l}$ to be J_{-l}, defined as:

$$J_{-l} = \frac{1}{N}||\mathbf{X}_{-l}\hat{\mathbf{p}}_{-l} - \mathbf{E}||_F^2 \tag{7.9}$$

where $\mathbf{X}_{-l}$ denotes the matrix $\mathbf{X}$ with the column corresponding to l-th feature removed. Then according to the definition, the utility of feature l, U_l is:

$$U_l = J_{-l} - J \tag{7.10}$$

A straightforward computation of U_l would be computationally heavy due to the fact that the computation of $\hat{\mathbf{p}}_l$ requires a matrix inversion of $\mathbf{X}_{-l}\mathbf{X}_{-l}^T$, which has to be repeated for each feature l.

However, it can be shown that the utility of the l-th feature of $\mathbf{X}$ in (7.10) can be computed efficiently without the explicit recomputation of $\hat{\mathbf{p}}_{-l}$ by using the following expression [18]:

$$U_l = \frac{1}{q_l}||\bar{\mathbf{p}}_l||_2, \tag{7.11}$$

where q_l is the l-th diagonal element of $\mathbf{R}_{\mathbf{XX}}^{-1}$ and p_l is the l-th row in $\hat{\mathbf{p}}$, corresponding to the l-th feature. The mathematical proof of (7.11) can be found in [18]. Note that $\mathbf{R}_{\mathbf{XX}}^{-1}$ is already known from the computation of $\hat{\mathbf{p}}$ such that no additional matrix inversion is required.

However, since the data matrix $\mathbf{X}$ can contain redundant features or features that are linear combinations of each other in its columns, it cannot be guaranteed that the matrix $\mathbf{X}$ in (7.7) is full-rank. In this case, the removal of a redundant column from $\mathbf{X}$ will not lead to an increase in the LS cost of (7.7). Moreover, $\mathbf{R}_{\mathbf{XX}}^{-1}$, used to find the solution of (7.7) in (7.8), will not exist in this case since the matrix $\mathbf{X}$ is rank deficient. A similar problem appears if $N < d$, which can happen in case of very high-dimensional data. To overcome this problem, the definition of utility generalized to a minimum l_2-norm selection [18] is used in this work. This approach eliminates the feature yielding the smallest increase in the l_2-norm of the weight matrix when the column corresponding to that feature were to be removed and the weight matrix would be re-optimized. Moreover, minimizing the l_2-norm of the weights further reduces the risk of overfitting.

This generalization is achieved by first adding an l_2-norm penalty β to the cost function that is minimized in (7.7):

$$J = \min_{\mathbf{p}} \frac{1}{2} ||\mathbf{X}\mathbf{p} - \mathbf{E}||_F^2 + \beta ||\mathbf{p}||_2^2 \tag{7.12}$$

where $0 < \beta \leqslant \mu$ with μ equal to the smallest non-zero eigenvalue of $\mathbf{R}_{\mathbf{XX}}$ in order to ensure that the bias added due to the penalty term in (7.12) is negligible. The minimizer of (7.12) is:

$$\hat{\mathbf{p}} = \mathbf{R}_{\mathbf{XX}\beta}^{-1} \mathbf{R}_{\mathbf{XE}} = (\mathbf{R}_{\mathbf{XX}} + \beta \mathbf{I})^{-1} \mathbf{R}_{\mathbf{XE}} \tag{7.13}$$

It is noted that (7.13) reduces to $\mathbf{R}_{\mathbf{XX}}^{\dagger} \mathbf{R}_{\mathbf{XE}}$ when $\beta \to 0$, where $\mathbf{R}_{\mathbf{XX}}^{\dagger}$ denotes the Moore-Penrose pseudo-inverse. This solution corresponds to the minimum norm solution of (7.7) when $\mathbf{X}$ contains linearly dependent columns or rows. The utility U_l of the l-th column in $\mathbf{X}$ based on (7.12) is [18]:

$$\begin{aligned} U_l &= \left(||\mathbf{X}_{-l} \hat{\mathbf{p}}_{-l} - \mathbf{E}||_2^2 - ||\mathbf{X}\hat{\mathbf{p}} - \mathbf{E}||_2^2 \right) \\ &\quad + \beta \left(||\hat{\mathbf{p}}_{-l}||_2^2 - ||\hat{\mathbf{p}}||_2^2 \right) \\ &= (J_{-l} - J) + \beta \left(||\hat{\mathbf{p}}_{-l}||_2^2 - ||\hat{\mathbf{p}}||_2^2 \right) \end{aligned} \tag{7.14}$$

Note that if column l in $\mathbf{X}$ is linearly independent from the other columns, (7.14) closely approximates to the original utility definition in (7.10) as the first term dominates over the second. However, if column l is linearly dependent, the first term vanishes and the second term will dominate. In this case, the utility quantifies the increase in l_2-norm after removing the l-th feature.

To select the best s features of $\mathbf{X}$, a greedy selection based on the iterative elimination of the features with the least utility is carried out. After the elimination of each feature, a re-estimation of the weights $\hat{\mathbf{p}}$ is carried out and the process of elimination is repeated, until s features remain.

Note that the value of β depends on the smallest non-zero eigenvalue of $\mathbf{R_{XX}}$. Since $\mathbf{R_{XX}}$ has to be recomputed every time when a feature is removed, also its eigenvalues change along the way. In practice, the value of β is selected only once and fixed for the remainder of the algorithm, as smaller than the smallest non-zero eigenvalue of $\mathbf{R_{XX}}$ before any of the features are eliminated [130]. This value of β will be smaller than all the non-zero eigenvalues of any principal submatrix of $\mathbf{R_{XX}}$ using the Cauchy's interlace theorem [81].

The summary of the utility subset selection is described in Algorithm 4. Algorithm 5 outlines the complete U2FS algorithm proposed in this work.

Algorithm 4 Utility metric algorithm for subset selection.

Input: Data $\mathbf{X}$, Eigenvectors $\mathbf{E}$, Number of features s to select
Output: s features selected

1: Calculate $\mathbf{R_{XX}}$ and $\mathbf{R_{XE}}$ as described in Equation (7.8).
2: Calculate β as the smallest non-zero eigenvalue of $\mathbf{R_{XX}}$
3: **while** Number of features remaining is $> s$ **do**
4: Compute $\mathbf{R}^{-1}_{\mathbf{XX}\beta}$ and $\hat{\mathbf{p}}$ as described in (7.13).
5: Calculate the utility of the remaining features using (7.11)
6: Remove the feature f_l with the lowest utility.
7: Update $\mathbf{R_{XX}}$ and $\mathbf{R_{XE}}$ by removing the rows and columns related to that feature f_l.
8: **end while**

Algorithm 5 Unsupervised feature selector based on the utility metric (U2FS).

Input: Data $\mathbf{X}$, Number of clusters c, Number of features s to select
Output: s features selected

1: Construct the similarity graph $\mathbf{W}$ as described in Section 7.3.1 selecting one of the weightings:
   - Binary
   - RBF kernel, using σ_0^2
   - RBF kernel, using $\hat{\sigma}^2$ based on Algorithm 3
2: Calculate the normalized Laplacian $\mathbf{L}$ and the eigenvectors α derived from Equation (7.2).
   Keep the c eigenvectors corresponding to the highest eigenvalues, excluding the first one.
3: Apply the backward greedy utility algorithm 4.
4: Return the s features remaining from the backward greedy utility approach.

As it has been stated before, one of the most remarkable aspects of the U2FS algorithm is the use of a greedy technique to solve the subset selection problem. The use of this type of method reduces the computational cost of the algorithm. This can be confirmed analyzing the computational complexity of U2FS, where the most demanding steps are the eigendecomposition of the Laplacian matrix (step 2 of Algortihm 5), which has a cost of $O(N^3)$ [173], and the subset selection stage in step 3 of Algorithm 5. Contrary to the state-of-the-art, the complexity of U2FS being a greedy method depends on the number of features to select. The most computationally expensive step of the subset selection in U2FS is the calculation of the matrix $\mathbf{R}_{\mathbf{XX}}^{-1}$, which has a computational cost of $O(d^3)$. In addition, this matrix needs to be updated $d-s$ times. This update can be done efficiently using a recursive updating equation from [18] with a cost of $O(t^2)$, with t the number of features remaining in the dataset at the current iteration of the greedy selection procedure, i.e. $t = d, d-1, d-2, ...s$. Since $t < d$, the cost for performing $d-s$ iterations will be $O((d-s)d^2)$, which depends on the number of features s to be selected. Note that the cost of computing the least squares solution $\hat{\mathbf{p}}_{-l}$ for each l in (7.14) is eliminated using the efficient equation (7.11), bringing down the cost for computing the utility from $O(t^4)$ to $O(t)$ in each iteration. This vanishes with respect to the $O(d^3)$ term (remember that $t < d$). Therefore, the total asymptotic complexity of U2FS is $O(N^3 + d^3)$.

7.4 Results

The aim of the following experiments is to evaluate the U2FS algorithm based on multiple criteria. With the focus on the new estimation of the embedding proposed, the

proposed RBF kernel approach using the estimated $\hat{\sigma}^2$ is compared to the σ_0^2 parameter proposed in [26], and to the binary KNN graph commonly used in [67]. On the other hand, the utility metric for subset selection is compared to other sparsity-inducing techniques, based on $l_p - norm$ regularizations. In these experiments, this is evaluated using the $l_1 - norm$. The outline of the different combinations considered in this work summarized in Table 7.1. The last method, $\text{RBF}_{\hat{\sigma}^2}$ + Utility, would be the one referred to as U2FS, combining the novelties suggested in this work.

Table 7.1: Methods compared in the experiments

	Similarity measure	**Subset selection**
$\mathbf{KNN}_{Bin}$ **+** $\mathbf{l_1 - norm}$	KNN + binary weighting	l_1-norm
$\mathbf{RBF}_{\sigma_0^2}$ **+** $\mathbf{l_1 - norm}$	RBF kernel, σ_0^2	l_1-norm
$\mathbf{KNN}_{Bin}$ **+ Utility**	KNN + binary weighting	Utility metric
$\mathbf{RBF}_{\sigma_0^2}$ **+ Utility**	RBF kernel, σ_0^2	Utility metric
$\mathbf{RBF}_{\hat{\sigma}^2}$ **+ Utility**	RBF kernel, $\hat{\sigma}^2$	Utility metric

These novelties are evaluated in two different scenarios, namely a simulation study, and in the application of the methods on benchmark datasets. In particular for the latter, the methods are not only evaluated in terms of accuracy, but also regarding computational cost. Additionally, U2FS is compared with 3 representative state-of-the-art algorithms from the general family of unsupervised sparsity-inducing feature selection algorithms:

- **MCFS**[26][1]. This algorithm served as inspiration to create U2FS, and therefore, it is added to the set of comparison algorithms as baseline reference. MCFS performs spectral embedding and l_1-norm regularization sequentially, and which served as inspiration to create U2FS.

- **NDFS**[101][2], which performs nonnegative spectral analysis with $l_{2,1}$-norm regularization. This algorithm is added to the experiments since it is an improvement of MCFS, while being the first algorithm simultaneously adapting both stages of manifold learning and subset selection. Therefore, NDFS represents the transition to these adaptive optimization-based feature selection algorithms.

- **RJGSC**[211] optimally derives the embedding of the data by adapting the results with $l_2, 1$-norm regularization. This algorithm is taken as a reference for the large class of adaptive sparsity-inducing feature selection algorithms, which are much more complex than U2FS, since they apply optimization to recursively adapt the embedding and feature selection stages of the methods. RJGSC was already

[1] http://www.cad.zju.edu.cn/home/dengcai/Data/MCFS.html
[2] http://www.cs.cmu.edu/ yiyang/Publications.html

compared to several feature selectors in [211], and therefore, it is taken here as upper-bound threshold in performance.

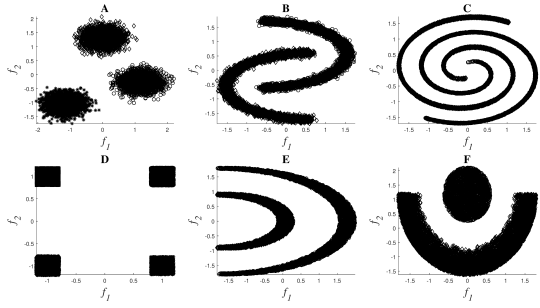

Figure 7.2: Toy examples used for simulations: Clouds (A), Moons (B), Spirals (C), Corners (D), Half-Kernel (E), Crescent Moon (F).

7.4.1 Simulations

A set of nonlinear toy examples typically used in clustering problems are proposed to test the different feature selection methods. In these experiments, the goal was to verify the correct selection of the original set of features. Figure 7.2 shows the toy examples considered[3], which are described by features f_1 and f_2 , and the final description of the datasets can be seen in Table 7.2.

All these problems are balanced, except for the last dataset Cres-Moon, for which the data is divided 25% to 75% between the two clusters. Five extra features in addition to the original f_1 and f_2 were added to each of the datasets in order to include redundant or irrelevant information:

- f_1' and f_2': random values extracted from two Pearson distributions characterized by the same higher-order statistics as f_1 and f_2 respectively.
- f_3' and f_4': Original f_1 and f_2 contaminated with Gaussian noise ($\nu\mathcal{N}(0,1)$), with $\nu = 1.5$.
- f_5': Constant feature of value 0.

The first step in the preprocessing of the features was to standardize the data using z-score to reduce the impact of differences in scaling and noise. In order to confirm

[3]The codes used to generate these datasets are available in https://github.com/avillago/u2fs

Table 7.2: Description of the toy example datasets.

	# samples	# classes
Clouds	9000	3
Moons	10000	2
Spirals	10000	2
Corners	10000	4
Half-Kernel	10000	2
Crescent-Moon	10000	2

the robustness of the feature selection techniques, the methods were applied using 10-fold cross-validation on the standardized data. For each fold a training set was selected using m-medoids, setting m to 2000 and using the centers of the clusters found as training samples. By doing so, the generalization ability of the methods can be guaranteed [178]. On each of the 10 training sets, the features were selected applying the 5 methods mentioned in Table 7.1. For each of the methods, the number of clusters c was introduced as the number of classes presented in Table 7.2. Since these experiments aim to evaluate the correct selection of the features, and the original features f_1 and f_2 are known, the number of features s to be selected was set to 2.

Regarding the parameter settings within the embedding methods, the binary was obtained setting k in the kNN approach to 5. For the RBF kernel embedding, σ_0^2 was set to the mean of the standard deviation along each dimension, as done in [26]. When using $\hat{\sigma}^2$, its value was obtained by applying the method described in Algorithm 3.

In terms of subset selection approaches, the method based on the $l_1 - norm$ automatically sets the value of the regularization parameter required for the LARS implementation, as described in [44]. For the utility metric, β was automatically set to the smallest non-zero eigenvalue of the matrix $\mathbf{R_{XX}}$ as described in Algorithm 4.

The performance of the algorithm is evaluated comparing the original set of features f_1 and f_2 to those selected by the algorithm. In these experiments, the evaluation of the selection results is binary: either the feature set selected is correct or not, regardless of the additional features f_i', for $i = 1, 2, ..., 5$, selected.

In Table 7.3 the most common results obtained in the 10 folds are shown. The utility-based approaches always obtained the same results for all 10 folds of the experiments. On the contrary, the $l_1 - norm$ methods provided different results for different folds of the experiment. For these cases, Table 7.3 shows the most common feature pair for each experiment, occurring at least 3 times.

As shown in Table 7.3, the methods that always obtain the adequate set of features are based on utility, both with the binary weighting and with the RBF kernel and the suggested $\hat{\sigma}^2$. Since these results were obtained for the 10 folds, they confirm both the

Table 7.3: Results feature selection for toy examples.

Method	**Utility metric**			$l_1 - norm$	
Embedding	**KNN**$_{Bin}$	**RBF**$_{\sigma_0^2}$	**RBF**$_{\hat{\sigma}^2}$	**KNN**$_{Bin}$	**RBF**$_{\sigma_0^2}$
Clouds	$\boldsymbol{f_1, f_2}$	f'_1, f'_4	$\boldsymbol{f_1, f_2}$	f'_1, f'_2	f'_1, f'_2
Moons	$\boldsymbol{f_1, f_2}$	f'_3, f'_4	$\boldsymbol{f_1, f_2}$	f'_1, f'_3	f'_1, f'_3
Spirals	$\boldsymbol{f_1, f_2}$	$\boldsymbol{f_1, f_2}$	$\boldsymbol{f_1, f_2}$	f_2, f'_2	f_2, f'_2
Corners	$\boldsymbol{f_1, f_2}$	f'_1, f'_2	$\boldsymbol{f_1, f_2}$	f_2, f'_2	f_2, f'_2
Half-Kernel	$\boldsymbol{f_1, f_2}$	f_2, f'_3	$\boldsymbol{f_1, f_2}$	f_1, f'_3	f_1, f'_3
Cres-Moon	$\boldsymbol{f_1, f_2}$	f_1, f'_4	$\boldsymbol{f_1, f_2}$	f_2, f'_1	f_2, f'_2

robustness and the consistency of the U2FS algorithm.

7.4.2 Benchmark datasets

Additionally, the proposed methods were evaluated using 6 well-known benchmark databases. The databases considered represent image (USPS, ORL, COIL20), audio (ISOLET) and text data (PCMAC, BASEHOCK)[4], proposing examples with more samples than features, and vice versa. The description of these databases is detailed in Table 7.4. All these datasets are balanced, except USPS.

Table 7.4: Description of the benchmark databases.

	Data Type	**Samples**	**Features**	**Classes**
USPS	Images	9298	256	10
Isolet	Audio	1560	617	26
ORL	Images	400	1024	40
COIL20	Images	1440	1024	20
PCMAC	Text	1943	3289	2
BASEHOCK	Text	1993	4862	2

In these datasets, the relevant features are unknown. Therefore, the common practice in the literature to evaluate feature selectors consists of applying the algorithms, taking from 10 to 80% of the original set of features, and evaluating the accuracy of a classifier when trained and evaluated with the selected feature set [211]. The classifier used for this aim in other studies is *k*-Nearest Neighbors (KNN), setting the number of neighbors to 5.

[4] All datasets downloaded from http://featureselection.asu.edu/datasets.php

These accuracy results are computed using 10-fold cross-validation to confirm the generalization capabilities of the algorithm. By setting m to 90% of the number of samples available in each benchmark dataset, m-medoids is used to select the m centroids of the clusters and use them as training set. Feature selection and the training of the KNN classifier are performed in these 9 folds of the standardized data, and the accuracy of the KNN is evaluated in the remaining 10% for testing. Exclusively for USPS, given the size of the dataset, 2000 samples were used for training and the remaining data was used for testing. These 2000 samples were also selected using m-medoids. Since PCMAC and BASEHOCK consist of binary data, these datasets were not standardized.

The parameters required for the binary and RBF embeddings, as well as β for the utility algorithm, are automatically set as detailed in section 7.4.1.

Figure 7.3 shows the median accuracy obtained for each of the 5 methods. The shadows along the lines correspond to the 25 and 75 percentile of the 10 folds. As a reference, the accuracy of the classifier without using feature selection is shown in black for each of the datasets. Additionally, Figure 7.4 shows the computation time for both the utility metric and the $l_1 - norm$ applied on a binary weighting embedding. In this manner, the subset selection techniques can be evaluated regardless of the code efficiency of the embedding stage. Similarly to Figure 7.3, the computation time plots show in bold the median running time for each of the subset selection techniques, and the 25 and 75 percentiles around it obtained from the 10-fold cross-validation.

The difference in the trends of the $l_1 - norm$ and utility in terms of computation time is due to their formulation. Feature selection based on $l_1 - norm$ regularization, solved using the LARS algorithm in this case, requires the same computation time regardless of the number of features aimed to select. All features are evaluated together, and later on, an MCFS score obtained from the regression problem is assigned to them [26]. The features with the higher scores are the ones selected. On the other hand, since the utility metric is applied in a backward greedy trend, the computation times change for different number of features selected. The lower the number of features selected compared to the original set, the higher the computation time. This is aligned with the computational complexity of the algorithm, described in Section 7.3.3. In spite of this, it can be seen that even the highest computation time for utility is lower than the time taken using $l_1 - norm$ regularization. The experiments were performed with 2x Intel Xeon E5-2640 @ 2.5 GHz processors and 64GB of working memory.

Finally, the experiments in benchmark databases are extended to compare U2FS to other key algorithms in the state-of-the-art. As it was mentioned at the beginning of this section, the selected algorithms are MCFS, NDFS, and RJGSC, which represent, respectively, the precursor of U2FS, an improved version of MCFS, and an example from the class of adaptive algorithms which recursively optimize the objective function proposed. NDFS and RJGSC require the tuning of their regularization

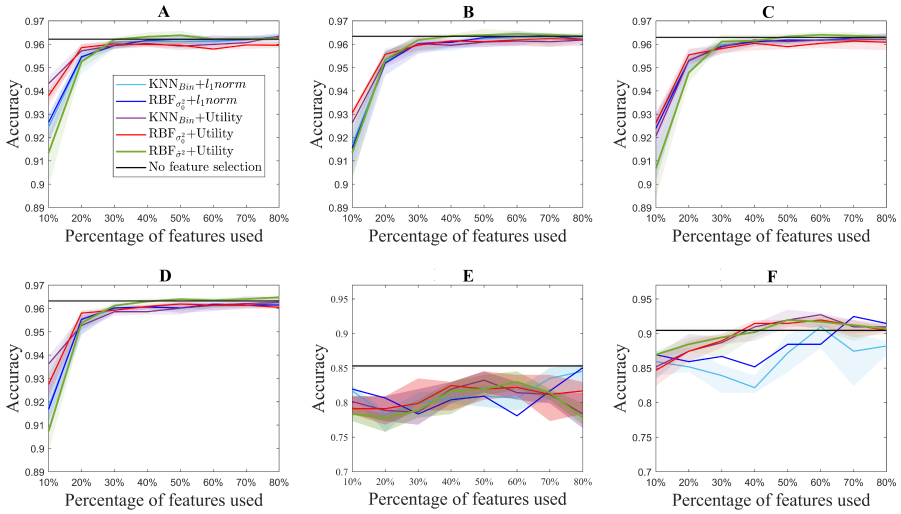

Figure 7.3:
Accuracy results for the benchmark databases, for selecting from 10 to 80% of the original number of features. The thick lines represent the median accuracy of the 10-fold cross-validation, and the shadows, the 25 and 75 percentile. USPS (Figure 3A), Isolet (B), ORL (C), COIL20 (D), PCMAC (E), BASEHOCK (F).

parameters, for which the indications in their corresponding articles were followed. For NDFS, the value of γ was set to 10^8, and α and β were selected from the values $\{10^{-6}, 10^{-4}, ..., 10^6\}$ applying grid search. The matrix F was initialized with the results of spectral clustering using all the features. For RJGSC, the results described in [211] for the BASEHOCK and PCMAC datasets are taken as a reference. In MCFS, the embedding is done using KNN and binary weighting, and the $l_1 - norm$ is used for subset selection. U2FS, on the other hand, results from the combination of the RBF kernel with $\hat{\sigma}^2$ and the utility metric. Table 7.5 summarizes the results by showing the KNN accuracy (ACC) for 10% of the features used, and the maximum ACC achieved among the percentages of features considered, for the BASEHOCK and PCMAC datasets.

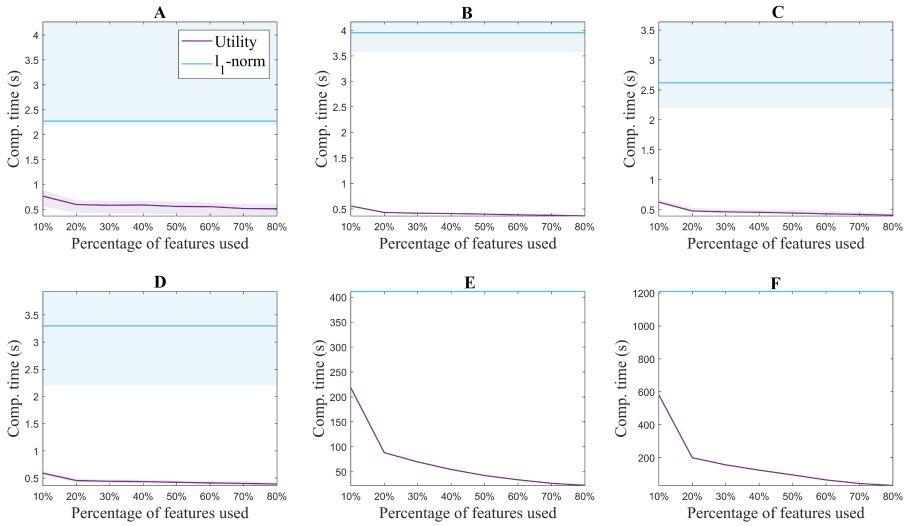

Figure 7.4:
Computation time for extracting from 10 to 80% of the original number of features for each of the benchmark databases. USPS (Figure 4A), Isolet (B), ORL (C), COIL20 (D), PCMAC (E), BASEHOCK (F).

Table 7.5: Comparison of classification accuracy (ACC) with the state-of-the-art for PCMAC and BASEHOCK datasets.

Dataset	**Method**	**ACC at 10% features**	**% features at Max ACC**	**Max ACC**
PCMAC	U2FS	0.785	60%	0.83
	MCFS	0.67	20%	0.697
	NDFS	0.73	40%	0.83
	RJGSC	0.805	60%	0.83
BASEHOCK	U2FS	0.87	50%	0.925
	MCFS	0.815	80%	0.84
	NDFS	0.76	20%	0.794
	RJGSC	0.902	80%	0.917

7.5 Discussion

The results obtained in the experiments suggest that the proposed U2FS algorithm obtains comparable results to the state-of-the-art in all the applications suggested, taking less computational time. Nevertheless, the performance of the utility metric for feature selection varies for the different experiments presented and requires a detailed analysis.

From Table 7.3, in Section 7.4.1, it can be concluded that the utility metric is able to select the correct features in an artificially contaminated dataset. Both the binary embedding and the RBF kernel with $\hat{\sigma}^2$ select the original set of features for the 10 folds of the experiment. The stability in the results also applies for the RBF embedding with σ_0^2, which always selected the same feature pair for all 10 folds even though they are only correct for the spirals problem.

Therefore, considering the stability of the results, it can be concluded that the proposed approach is more robust in the selection of results than that based on the $l_1 - norm$.

On the other hand, when considering the suitability of the features selected, two observations can be made. First of all, it can be seen that the lack of consistency in the $l_1 - norm$ approaches discards the selection of the correct set of features. Moreover, the wrong results obtained with both $l_1 - norm$ and utility methods for the RBF embedding using σ_0^2 reveal the drawback of applying this approximation of σ_0^2 in presence of redundant or irrelevant features. Since this value is calculated as the mean of the standard deviation of all the dimensions in the data, this measure can be strongly affected by irrelevant data, that could be very noisy and enlarge this sigma, leading to the allocation of all the samples to a mega-cluster.

While the use of the proposed approximation for $\hat{\sigma}^2$ achieves better results than σ_0^2, these are comparable to the ones obtained with the KNN binary embedding when using the utility metric. The use of KNN to build graphs is a well-known practice, very robust for dense clusters, as it is the case in these examples. The definition of a specific field where each of the embeddings would be superior is beyond the scope of this study. However, the excellence of both methods when combined with the proposed subset selection method only confirms the robustness of the utility metric, irrespective of the embedding considered.

For standardization purposes, the performance of the method was evaluated in benchmark databases. As it can be observed, in terms of the accuracy obtained for each experiment, U2FS achieves comparable results to the $l_1 - norm$ methods for most of the datasets considered, despite its condition of greedy method.

In spite of this, some differences in performance can be observed in the different datasets. The different ranking of the methods, as well as the accuracy obtained for

each of the databases can be explained taking into account the type of data under study and the ratio between samples and dimensions.

With regard to the type of data represented by each test, it can be observed that for the ISOLET dataset, containing sound information, two groups of results are distinguishable. The group of the utility metric results outperforms those derived from the $l_1 - norm$, which only reach comparable results for 60% of the features selected. These two groups of results are caused by the subset selection method applied, and not for the embedding, among which the differences are not remarkable.

In a similar way, for the case of the image datasets USPS, ORL and COIL20, the results derived from utility are slightly better than those coming from the $l_1 - norm$. In these datasets, similarly to the performance observed in ISOLET, accuracy increases with the number of features selected.

Regarding the differences between the proposed embeddings, it can be observed that the results obtained are comparable for all of them. Nonetheless, Figure 7.3 shows that there is a slight improvement in the aforementioned datasets for the RBF kernel with $\hat{\sigma}^2$, but the results are still comparable to those obtained with other embeddings. Moreover, this similarity in the binary and RBF results holds for the $l_1 - norm$ methods, for which the accuracy results almost overlap in Figure 7.3. This can be explained by the relation between the features considered. Since for these datasets the samples correspond to pixels, and the features to the color codes, a simple neighboring method such as the binary weighting is able to code the connectivity of pixels of similar colors.

The text datasets, PCMAC and BASEHOCK, are the ones that show bigger differences between the results obtained with utility and those obtained with the $l_1 - norm$. This can be explained by the amount of zeros present in the data, with which the utility metric is able to cope slightly better. The sparsity of the data leads to more error in the $l_1 - norm$ results, since more features end up having the same MCFS score, and among those, the order for selection comes at random. The results obtained with the utility metric are more stable, in particular for the BASEHOCK dataset. For this dataset, U2FS even outperforms the results without feature selection if at least 40% of the features are kept.

In all the datasets proposed, the results obtained with the $l_1 - norm$ show greater variability, i.e. larger percentiles. This is aligned with the results obtained in the simulations. The results for the $l_1 - norm$ are not necessarily reproducible in different runs, since the algorithm is too sensitive to the training set selected. The variability of the utility methods is greater for the approaches based on the RBF kernel. This is due to the selection of the σ^2 parameter, which also depends on the training set. The tuning of this parameter is still very sensitive to high-dimensional and large-scale data, posing a continuous challenge for the machine learning community [199, 171].

Despite it being a greedy method, the utility metric proves to be applicable to feature

selection approaches and to strongly outperform the $l_1 - norm$ in terms of computational time, without significant reduction in accuracy. U2FS proves to be effective both in cases with more samples than features and vice versa. The reduction in computation time is clear, for all the benchmark databases described, and is particularly attractive for high-dimensional datasets. Altogether, our feature selection approach U2FS, based on the utility metric, and with the binary or the RBF kernel with $\hat{\sigma}^2$ is recommended due to its fast performance and its interpretability.

Additionally, the performance of U2FS is comparable to the state-of-the-art, as shown in Table 7.5. In this table, the performance of U2FS (RBF kernel and $\hat{\sigma}^2$, with the utility metric) is compared to that of MCFS, NDFS and RJGSC. For MCFS, it can be seen that, as expected, U2FS appears as an improvement of this algorithm, achieving better results for both datasets. For NDFS, the results are slightly worse than for U2FS, most probably due to problems in the tuning of regularization parameters. Given the consistent good results for different datasets of RJGSC when compared against the state-of-the-art, and its condition of simultaneously adapting the spectral embedding and subset selection stages, this algorithm is taken as example of the most complex SSFS algorithms (SAMM-FS, SOGFS or DSRMR). These algorithms perform manifold learning and feature selection simultaneously, iteratively adapting both steps to achieve optimal results.

It is clear that in terms of accuracy, both for 10% of the features and for the maximal value of achieved, U2FS obtains similar results to RJGSC, while at the same time having a much smaller computational complexity. Furthermore, while RJGSC requires the manual tuning of extra parameters, similarly to other algorithms in the state-of-the-art, U2FS tunes its parameters automatically. Hence, the application of the method is straightforward for the users. The stages of higher complexity in U2FS, previously defined as $O(N^3 + d^3)$, are shared by most of the algorithms in the state-of-the-art. However, on top of these eigendecompositions and matrix inversions, the algorithms in the literature require a number of iterations in the optimization process that U2FS avoids. Additionally, U2FS is the only algorithm for which the computation time scales linearly with the amount of features selected.

The current state-of-the-art of unsupervised spectral feature selectors applies the stages of manifold learning and subset selection simultaneously, which can lead to optimal results. In a field that gets more and more complex and goes far from applicability, U2FS is presented as a quick solution for a sequential implementation of both stages of SSFS algorithms, yet achieving comparable results to the state-of-the-art. Being a greedy method, the utility metric cannot be applied simultaneously to the manifold learning and subset selection stages. However, other sequential algorithms from the state-of-the-art could consider the use of utility for subset selection, instead of the current sparsity-inducing techniques. One of the most direct applications could be the substitution of group-LASSO for group-utility, in order to perform selections of groups of features as proposed by [18]. This can be of interest in cases where the relations

between features are known, such as in channel selection [130] or in multi-modal applications [209].

7.6 Conclusion

This work presents a new method for unsupervised feature selection based on manifold learning and sparse regression. The main contribution of this chapter is the formulation of the utility metric in the field of spectral feature selection, substituting other sparse regression methods that require more computational resources. This method, being a backward greedy approach, has been proven to obtain comparable results to the state-of-the-art methods with analogous embedding approaches, yet at considerably reduced computational load. The method shows consistently good results in different applications, from images to text and sound data; and it is broadly applicable to problems of any size: using more features than samples or vice versa.

Furthermore, aiming to show the applicability of U2FS to data presenting non-linearities, the proposed approach has been evaluated in simulated data, considering both a binary and an RBF kernel embedding. Given the sensitivity of the RBF kernel to high-dimensional spaces, a new approximation of the RBF kernel parameter was proposed, which does not require further tuning around the value obtained. The proposed approximation outperforms the rule-of-thumb widely used in the literature in most of the scenarios presented. Nevertheless, in terms of feature selection, the utility metric is robust against the embedding.

U2FS is proposed as a non-parametric efficient algorithm, which does not require any manual tuning or special knowledge from the user. Its simplicity, robustness and accuracy open a new path for structure sparsity-inducing feature selection methods, which can benefit from this quick and efficient technique.

Chapter 8

Conclusions and future work

8.1 Conclusions

Multi-lead ECG analysis is key in the diagnosis of arrhythmogenic cardiac disease. These ECG signals are recorded in routine clinical examination in search of abnormal patterns that can suggest any risk for arrhythmia. Several risk markers have been proposed in the literature, and most of them can be automatically detected. Multi-lead ECG provides a graphical representation of the electrical cardiac cycle captured from different angles. While the visual analysis of ECG signals by clinicians takes into account the information contained in the different leads, most of the automated approaches in the literature either treat each lead independently, or only evaluate one of the leads.

Therefore, the goal of this thesis was to propose advanced high-dimensional techniques for the analysis of 12-lead ECG signals. This goal was divided in three research objectives:

1. Develop robust algorithms for the extraction of arrhythmia risk predictors from the ECG.
2. Propose a novel framework for dimensionality reduction of 12-lead ECG signals that enables the comparison of different subjects and enhances irregularities.
3. Develop an unsupervised feature selector based on manifold learning.

8.1.1 Robust algorithms for the extraction of arrhythmia risk predictors from the ECG

The first goal of this thesis was to develop robust algorithms to extract arrhythmia risk predictors from the ECG. Chapters 2-6 contributed to this goal by proposing novel algorithms which were compared against the state-of-the-art.

Chapter 2 presented an algorithm for heartbeat classification in single-lead ECG signals. The method was designed to be used on signals recorded from wearable devices, exploiting the temporal and frequency content of single-lead ECG while being robust against noise. Its performance for the most common heartbeat classes normal (N), supraventricular and ventricular ectopic beats (SVEB and VEB) was comparable to the state-of-the-art. Nevertheless, the results presented some limitations in the identification of fusion beats (F). Additionally, all the methods considered in the evaluation of this work performed poorly for the SVEB class. These generalized results triggered an analysis of the training and testing protocol considered for the evaluation of these algorithms. This study concluded that the database commonly used in the literature is too limited and does not offer enough beats of the SVEB and F classes, which affects the performance of any algorithm.

Chapter 3 presents a study exploring the relation of fQRS with an increased risk of arrhythmia in ICD patients, by exploring the case of patients presenting with electrical storm (ES). This study encountered some challenges that motivated the rest of the chapters related to this research objective. First, the algorithms used for QRS segmentation and fQRS quantification had several limitations [64]. The visual inspection of the results highlighted errors in the segmentation, which motivated the development and publication of the multi-lead heartbeat classification algorithm presented in Chapter 4. Additionally, the fQRS quantification approach was exclusively trained on a single database, and on a single definition of fragmentation. Therefore, a new algorithm for fQRS quantification was presented in Chapter 4, which was trained and evaluated on multi-center data. This algorithm obtained more accurate results than the algorithm proposed in [64] thanks to the segmentation approach, and its performance was evaluated on data from different centers. It was concluded that the best model according to most of the performance metrics was the one trained on multi-center data, given its generalization capabilities and its performance on both datasets included in this study.

Chapters 5 and 6 explored the use of novel and state-of-the-art arrhythmia risk markers by comparing the ECGs of ICD patients with and without history of ES. None of these features could identify differences between the ECGs of these two groups of patients, but they could distinguish between these two groups and the ECGs of healthy subjects matched based on sex and age. The database for the ES study had clear limitations that affected the characterization of both groups of patients. On the one hand, both

the ES and non-ES groups contained IHD and DCM patients, which were shown to present differences in the absence of arrhythmogenic events. Additionally, the size of the dataset was very limited and the time between the recording of the ECG for the different visits and the ES event differed considerably between the different patients.

8.1.2 Framework for dimensionality reduction of 12-lead ECG signals

The analysis of the state-of-the-art in Section 1.3 stressed the lack of methods that integrate the information in 12-lead ECG to derive arrhythmia risk predictors. Therefore, the second goal of this thesis was to propose a framework to reduce the dimensionality of 12-lead ECG signals. Chapter 5 presents this framework, which uses Laplacian Eigenmaps for the definition of a reference 3-D space created from normal 12-lead ECGs, onto which new ECGs can be mapped. Laplacian Eigenmaps are a data-driven dimensionality reduction approach that allows to translate a heartbeat in 12 dimensions into a 3-D loop that captures the irregularities of the ECG. This framework included the definition of a set of distance metrics and features extracted from these loops that allowed the comparison of different patients. The method was proven to capture the characteristics of normal ECGs, and to differentiate these from patients with ischemic and non-ischemic cardiomyopathy.

Additionally, the added value of using a data-driven dimensionality reduction approach was emphasized by comparing the framework proposed in Chapter 5 to VCG. Since it can be mathematically derived from 12-lead ECG, VCG can be considered a dimensionality reduction method. However, the transformation is similar to any ECG signal, regardless of its characteristics. Hence, VCGs derived from 12-lead may remove information that is kept using Laplacian Eigenmaps, while the latter is able to remove irrelevant information or noise. This was concluded given the variability of the results derived from the VCG approach, even for different groups of normal ECGs.

8.1.3 Unsupervised feature selector based on manifold learning

The last objective of this thesis was to develop an unsupervised feature selector based on manifold learning. This was achieved by proposing the U2FS algorithm presented in Chapter 7, which was proven to perform comparably to the state-of-the-art while offering several advantages. The motivation to pursue this goal came from the lack of publicly available algorithms for unsupervised feature selection. Additionally, most of the methods proposed in the literature implied the tuning of several parameters, making their implementation problematic.

The U2FS algorithm tackles these problems combining manifold learning for the stage of exploration and identification of relevant structures in high-dimensional data, and the utility metric for subset selection. Its first advantage is that it only requires as inputs the number of features to be selected and the number of expected clusters in the data, making it user-friendly for its application in any type of data. Second, the method is applicable to data presenting non-linearities, by using an RBF kernel in the embedding stage. One of the additional contributions of this work is the development of an algorithm for the automatic tuning of the RBF kernel parameter, which strongly affects the results of the feature selector. The proposed approach was proven to perform better than the rule of thumb commonly used in clustering approaches, leading to consistent results with a much simpler methodology. Third, the use of the utility metric for subset selection makes U2FS less computationally expensive and faster than other sparse regression techniques, while achieving comparable results. Finally, this user-friendly and efficient unsupervised feature selector was made publicly available for its application in any type of high-dimensional data.

Despite U2FS being a powerful tool, it was not possible to demonstrate its applicability in the problems tackled in this thesis. The reason for this is that it remains unsupervised, and hence the connection between the findings and any clinical outcome is still a matter of research. Future steps to bridge this gap will be discussed in Section 8.2.3.

8.2 Future work

8.2.1 Robust algorithms for the extraction of arrhythmia risk predictors from the ECG

Chapter 2 showed the impact of the patient selection for training and test in heartbeat classification algorithms. This is a requirement for several biomedical applications, where different patients are included in the training and testing to avoid bias in the algorithms. The splitting of patients in different study groups is mostly done manually and it requires the evaluation of several characteristics to guarantee a balanced and matched distribution of the patients. Future work should explore automated methods for this task, which is crucial for machine learning approaches applied to clinical data.

About the analysis of fQRS, some open questions remain that should be addressed in future work. Chapter 4 presented a novel algorithm for the quantification of fQRS. Despite of its improvements related to the training and testing in multi-center data, the knowledge of this algorithm is limited to the definitions included on the considered datasets. The analysis of fQRS lacks a gold standard that unifies the multiple definitions and reduces the current inter and intra-observer variability. Future work should include other fQRS definitions in the training and evaluation of these automated

approaches, with the aim of creating an objective metric that can support clinical diagnosis. Additionally, the output of the proposed model is computed separately for each of the 12-leads. Accordingly, the joint interpretation of those 12-independent fQRS scores still involves the analysis of high-dimensional data with its corresponding challenges. Future work should explore the development of a global fQRS metric, which would synthesize the scores in each of the leads. Additionally, the definition of regional scores which could characterize the impact of fragmentation in the different regions of the heart should be explored.

Chapters 3, 5 and 6 aimed to identify differences between ICD patients with and without history of ES. However, while the first study identified a trend of increasing average level of fQRS with the increasing risk of ES, there were no significant differences in the level of fQRS between ES patients and controls for the ECGs recorded around the time of the event. Similarly, the studies presented in Chapters 5 and 6 could neither identify differences between these two populations using VCG nor the LE framework. Future work should continue exploring these differences, which will require a larger database that would allow the separation of IHD and DCM patients, and a more uniform time difference between the visits for the analysis of changes over time.

In terms of methodology, the methods proposed in Chapters 5 and 6 were only applied to the stage of ventricular depolarization of the ECG. This was the case in order to continue with the analysis related to fQRS, which also refers to this period, for the characterization of structural or scar-related abnormalities in the heart. Additionally, the analysis of ventricular depolarization reduces the impact of segmentation errors on the evaluation of LE and VCG, due to the simpler delineation of the QRS-complex compared to the T-wave. Nevertheless, future work should explore the extension of the LE and VCG approaches to the analysis of ventricular repolarization, which is related to inherent cellular electrophysiological changes [79, 92] and has been thoroughly studied for arrhythmia prediction [11, 70, 200].

8.2.2 Framework for dimensionality reduction of 12-lead ECG signals

The relevance of the LE framework proposed in Chapter 5 was confirmed by its results on normal ECGs and signals from ischemic and non-ischemic cardiomyopathy patients. The reference space was built using ECGs of male and female subjects with ages ranged between 50 and 75 years to much the patient populations under study. Future studies should explore the behavior of LE in men and women, and for different ranges of age.

Additionally, both the LE framework and VCG were treated as new sources of features for patient characterization. While distance metrics were proposed to analyze the

complete LE loops in Chapter 5, these distances were simplified to a single metric, which did not explore changes along the loop. Future work should explore the comparison of the entire loops between patient populations, both for VCG and LE, for a less restricted analysis of signal abnormalities.

Finally, Chapter 6 offered a first comparison between VCG and LE as ECG dimensionality reduction approaches. Future work should compare them in more detail, by using bigger datasets and exploring the limitations of each of these approaches, and explore the potential of combining both of them. Moreover, other dimensionality reduction techniques like SVD or tensors should be compared to VCG and LE, providing recommendations for their optimal application in multi-lead ECG analysis.

8.2.3 Unsupervised feature selector based on manifold learning

Chapter 7 proposes an unsupervised feature selector based on manifold learning and the utility metric. The latter has been proven to reduce the complexity and the computation time of the subset selection stage of the algorithm. Therefore, future work should explore the use of this method to substitute other sparse regression approaches. One example of this could be changing group-LASSO for group-utility, for the selection of groups of features already predefined.

The U2FS algorithm only requires as input the number of clusters and the number of features to be kept. However, these values are not always known, and hence their value is assigned by the user by trial and error. Future studies should consider the automatic definition of these inputs, by optimizing clustering performance metrics.

Even though the work on U2FS was inspired by the study on heartbeat classification, this algorithm was not applied in any of the other chapters of this thesis. This is due to the challenges related to the use of unsupervised approaches, in particular in terms of interpretability. This becomes even more relevant in the clinical field, where the analysis of results should lead to medical conclusions. Therefore, future work should aim to bring these powerful methods closer to clinical applications, which is a task that requires the close collaboration between clinicians and data scientists.

Appendix A

Appendix

A.1 Supplementary material for Chapter 5

A.1.1 Table for results of different kernels, Experiment 1

Table A.1 presents the results for Experiment 1 related to Section 4.3.1, comparing the use of the linear, polynomial and RBF kernel in the SVM classifier.

Table A.1:

Results for Experiment 1 for each of the 3 kernels considered in the SVM classifier. The results are expressed as mean ± standard deviation, summarizing the 10 repetitions performed per kernel.

Approach	Kernel	Sens	Spec	PPV	ROC AUC	PR AUC	κ
Proposed	Linear	0.74 ± 0.02	0.92 ± 0.01	0.84 ± 0.02	0.92 ± 0.01	0.86 ± 0.02	0.69 ± 0.02
	Polynomial	0.70 ± 0.09	0.92 ± 0.01	0.84 ± 0.03	0.91 ± 0.03	0.85 ± 0.04	0.67 ± 0.06
	RBF	0.73 ± 0.02	0.92 ± 0.01	0.84 ± 0.02	0.92 ± 0.01	0.87 ± 0.02	0.69 ± 0.02
Reference	Linear	0.67 ± 0.03	0.89 ± 0.02	0.78 ± 0.02	0.88 ± 0.01	0.82 ± 0.02	0.62 ± 0.02
	Polynomial	0.67 ± 0.07	0.90 ± 0.02	0.80 ± 0.02	0.89 ± 0.02	0.83 ± 0.02	0.62 ± 0.04
	RBF	0.69 ± 0.04	0.90 ± 0.02	0.81 ± 0.02	0.88 ± 0.01	0.83 ± 0.02	0.63 ± 0.02

Bibliography

[1] ACAR, B., AND KOYMEN, H. SVD-based on-line exercise ECG signal orthogonalization. *IEEE Transactions on Biomedical Engineering 46*, 3 (1999), 311–321.

[2] AGGARWAL, C. C., HINNEBURG, A., AND KEIM, D. A. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory* (2001), Springer, pp. 420–434.

[3] AHMAD, M., KHAN, A. M., MAZZARA, M., DISTEFANO, S., ALI, A., AND TUFAIL, A. Extended sammon projection and wavelet kernel extreme learning machine for gait-based legitimate user identification. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (2019), pp. 1216–1219.

[4] ALZATE, C., AND SUYKENS, J. A. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. *IEEE transactions on pattern analysis and machine intelligence 32*, 2 (2008), 335–347.

[5] ANZAI, Y. *Pattern recognition and machine learning*. Elsevier, 2012.

[6] ARO, A. L., ANTTONEN, O., TIKKANEN, J. T., JUNTTILA, M. J., KEROLA, T., RISSANEN, H. A., REUNANEN, A., AND HUIKURI, H. V. Intraventricular conduction delay in a standard 12-lead electrocardiogram as a predictor of mortality in the general population. *Circulation: Arrhythmia and Electrophysiology 4*, 5 (oct 2011), 704–710.

[7] ARO, A. L., HUIKURI, H. V., TIKKANEN, J. T., JUNTTILA, M. J., RISSANEN, H. A., REUNANEN, A., AND ANTTONEN, O. QRS-T angle as a predictor of sudden cardiac death in a middle-aged general population. *Europace 14*, 6 (jun 2012), 872–876.

[8] ARONIS, K. N., ALI, R. L., PRAKOSA, A., ASHIKAGA, H., BERGER, R. D., HAKIM, J. B., LIANG, J., TANDRI, H., TENG, F., CHRISPIN, J., AND TRAYANOVA, N. A. Accurate conduction velocity maps and their

association with scar distribution on Magnetic Resonance Imaging in patients with postinfarction ventricular tachycardias. *Circulation: Arrhythmia and Electrophysiology 13*, 4 (apr 2020), 326–335.

[9] ARYA, A., HAGHJOO, M., DEHGHANI, M. R., FAZELIFAR, A. F., NIKOO, M.-H., BAGHERZADEH, A., AND SADR-AMELI, M. A. Prevalence and predictors of electrical storm in patients with implantable cardioverter-defibrillator. *The American journal of cardiology 97*, 3 (2006), 389–392.

[10] ASSOCIATION FOR THE ADVANCEMENT OF MEDICAL INSTRUMENTATION AND OTHERS. Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms. *ANSI/AAMI EC38* (1998).

[11] ATIGA WL, CALKINS H, LAWRENCE JH, TOMASELLI GF, SMITH JM, B. R. Beat-to-Beat repolarization lability identifies patients at risk for sudden cardiac death. *Journal of Cardiovascular Electrophysiology 9*, 9 (sep 1998), 899–908.

[12] BASTIAENEN, R., BATCHVAROV, V., AND GALLAGHER, M. M. Ventricular automaticity as a predictor of sudden death in ischaemic heart disease. *Europace 14*, 6 (2012), 795–803.

[13] BAUER, A., KANTELHARDT, J. W., BUNDE, A., BARTHEL, P., SCHNEIDER, R., MALIK, M., AND SCHMIDT, G. Phase-rectified signal averaging detects quasi-periodicities in non-stationary data. *Physica A: Statistical Mechanics and its Applications 364* (2006), 423–434.

[14] BAUER, A., MALIK, M., SCHMIDT, G., BARTHEL, P., BONNEMEIER, H., CYGANKIEWICZ, I., GUZIK, P., LOMBARDI, F., MÜLLER, A., OTO, A., ET AL. Heart rate turbulence: standards of measurement, physiological interpretation, and clinical use: International society for holter and noninvasive electrophysiology consensus. *Journal of the American College of Cardiology 52*, 17 (2008), 1353–1365.

[15] BELKIN, M., AND NIYOGI, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems* (2002), pp. 585–591.

[16] BELKIN, M., AND NIYOGI, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation 15*, 6 (2003), 1373–1396.

[17] BERAZA, I., AND ROMERO, I. Comparative study of algorithms for ECG segmentation. *Biomedical Signal Processing and Control 34* (2017), 166–173.

[18] BERTRAND, A. Utility metrics for assessment and subset selection of input variables for linear estimation [tips & tricks]. *IEEE Signal Processing Magazine 35*, 6 (2018), 93–99.

[19] BIGGS, N., BIGGS, N. L., AND NORMAN, B. *Algebraic graph theory*, vol. 67. Cambridge university press, 1993.

[20] BINICI, Z., INTZILAKIS, T., NIELSEN, O. W., KØBER, L., AND SAJADIEH, A. Excessive supraventricular ectopic activity and increased risk of atrial fibrillation and stroke. *Circulation 121*, 17 (2010), 1904–1911.

[21] BONO, V., MAZOMENOS, E. B., CHEN, T., ROSENGARTEN, J. A., ACHARYYA, A., MAHARATNA, K., MORGAN, J. M., AND CURZEN, N. Development of an automated updated selvester QRS scoring system using SWT-based QRS fractionation detection and classification. *IEEE journal of biomedical and health informatics 18*, 1 (2013), 193–204.

[22] BONOW, R. O., MANN, D. L., ZIPES, D. P., AND LIBBY, P. *Braunwald's heart disease e-book: A textbook of cardiovascular medicine*. Elsevier Health Sciences, 2011.

[23] BORTOLAN, G., CHRISTOV, I., SIMOVA, I., AND DOTSINSKY, I. Noise processing in exercise ECG stress test for the analysis and the clinical characterization of QRS and T wave alternans. *Biomedical Signal Processing and Control 18* (apr 2015), 378–385.

[24] BRIERS, T. Exploring the potential of vectorcardiography for prediction of ventricular storm in icd patients. Master's thesis, KU Leuven. Faculteit Ingenieurswetenschappen, 2021.

[25] BURCH, G. E. The history of vectorcardiography. *Medical History 29*, S5 (1985), 103–131.

[26] CAI, D., ZHANG, C., AND HE, X. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (2010), pp. 333–342.

[27] CAMM, A. J., MALIK, M., BIGGER, J. T., BREITHARDT, G., CERUTTI, S., COHEN, R., COUMEL, P., FALLEN, E., KENNEDY, H., KLEIGER, R., ET AL. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology.

[28] CASTRO, I., VARON, C., MOEYERSONS, J., GOMEZ, A. V., MORALES, J., DEVIAENE, M., TORFS, T., VAN HUFFEL, S., PUERS, R., AND VAN HOOF, C. Data quality assessment of capacitively-coupled ecg signals. In *2019 Computing in Cardiology (CinC)* (2019), IEEE, pp. Page–1.

[29] CHEN, X., HANSSON, P.-O., THUNSTRÖM, E., MANDALENAKIS, Z., CAIDAHL, K., AND FU, M. Incremental changes in QRS duration as predictor

for cardiovascular disease: a 21-year follow-up of a randomly selected general population. *Scientific Reports 11*, 1 (2021), 1–6.

[30] CHUNG, F. R., AND GRAHAM, F. C. *Spectral graph theory*. No. 92. American Mathematical Soc., 1997.

[31] CLIFFORD, G. D., AZUAJE, F., MCSHARRY, P. E., BAILON, R., SORNMO, L., AND LAGUNA, P. *Advanced Methods and Tools for ECG Data Analysis*, vol. 1. Boston: Artech house, 2006.

[32] CORREA, R., ARINI, P. D., VALENTINUZZI, M. E., AND LACIAR, E. Novel set of vectorcardiographic parameters for the identification of ischemic patients. *Medical Engineering and Physics 35*, 1 (jan 2013), 16–22.

[33] CORTES, C., AND VAPNIK, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[34] CORTEZ, D. L., AND SCHLEGEL, T. T. When deriving the spatial QRS-T angle from the 12-lead electrocardiogram, which transform is more frank: regression or inverse dower? *Journal of Electrocardiology 43*, 4 (2010), 302–309.

[35] COUVREUR, C., AND BRESLER, Y. On the optimality of the backward greedy algorithm for the subset selection problem. *SIAM Journal on Matrix Analysis and Applications 21*, 3 (2000), 797–808.

[36] DAS, M. K., KHAN, B., JACOB, S., KUMAR, A., AND MAHENTHIRAN, J. Significance of a fragmented QRS complex versus a Q wave in patients with coronary artery disease. *Circulation 113*, 21 (may 2006), 2495–2501.

[37] DAS, M. K., MASKOUN, W., SHEN, C., MICHAEL, M. A., SURADI, H., DESAI, M., SUBBARAO, R., AND BHAKTA, D. Fragmented QRS on 12-lead electrocardiogram predicts arrhythmic events in patients with ischemic and nonischemic cardiomyopathy. *Heart rhythm 7*, 1 (2010), 74–80.

[38] DAS, M. K., SURADI, H., MASKOUN, W., MICHAEL, M. A., SHEN, C., PENG, J., DANDAMUDI, G., AND MAHENTHIRAN, J. Fragmented wide QRS on a 12-lead ECG: a sign of myocardial scar and poor prognosis. *Circulation: Arrhythmia and Electrophysiology 1*, 4 (2008), 258–268.

[39] DAS, M. K., AND ZIPES, D. P. Fragmented qrs: a predictor of mortality and sudden cardiac death. *Heart Rhythm 6*, 3 (2009), S8–S14.

[40] DAWSON, D., YANG, H., MALSHE, M., BUKKAPATNAM, S. T., BENJAMIN, B., AND KOMANDURI, R. Linear affine transformations between 3-lead (Frank XYZ leads) vectorcardiogram and 12-lead electrocardiogram signals. *Journal of Electrocardiology 42*, 6 (2009), 622–630.

[41] De Chazal, P., O'Dwyer, M., and Reilly, R. B. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE transactions on biomedical engineering 51*, 7 (2004), 1196–1206.

[42] de Luna, A. B., and Fiol-Sala, M. *Electrocardiography in ischemic heart disease: clinical and imaging correlations and prognostic implications*. John Wiley & Sons, 2008.

[43] Demski, A., and Soria, M. L. ecg-kit: a Matlab toolbox for cardiovascular signal processing. *Journal of open research software 4*, 1 (2016).

[44] Deng Cai, Chiyuan Zhang, X. H. Supervised/Unsupervised/Semi-supervised Feature Selection for Multi-Cluster/Class Data, 2020.

[45] Deviaene, M., Testelmans, D., Borzee, P., Buyse, B., Van Huffel, S., and Varon, C. Feature selection algorithm based on random forest applied to sleep apnea detection. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2019), IEEE, pp. 2580–2583.

[46] Dhar, R., Alsheikh-Ali, A. A., Estes, N. A., Moss, A. J., Zareba, W., Daubert, J. P., Greenberg, H., Case, R. B., and Kent, D. M. Association of prolonged QRS duration with ventricular tachyarrhythmias and sudden cardiac death in the Multicenter Automatic Defibrillator Implantation Trial II (MADIT-II). *Heart Rhythm 5*, 6 (jun 2008), 807–813.

[47] Dorian, P., Hohnloser, S. H., Thorpe, K. E., Roberts, R. S., Kuck, K. H., Gent, M., and Connolly, S. J. Mechanisms underlying the lack of effect of implantable cardioverter-defibrillator therapy on mortality in high-risk patients with recent myocardial infarction: Insights from the Defibrillation in Acute Myocardial Infarction Trial (DINAMIT). *Circulation 122*, 25 (2010), 2645–2652.

[48] Dragomiretskiy, K., and Zosso, D. Variational mode decomposition. *IEEE transactions on signal processing 62*, 3 (2013), 531–544.

[49] Edenbrandt, L., and Pahlm, O. Vectorcardiogram synthesized from a 12-lead ECG: Superiority of the inverse Dower matrix. *Journal of Electrocardiology 21*, 4 (jan 1988), 361–367.

[50] Eifling, M., et al. The evaluation and management of electrical storm. *Texas Heart Institute Journal 38*, 2 (2011), 111.

[51] Ellenbogen, K. A., Wilkoff, B. L., Kay, G. N., and Lau, C. P. *Clinical Cardiac Pacing, Defibrillation and Resynchronization Therapy E-Book*. Elsevier Health Sciences, 2011.

[52] ENGELSTEIN, E. Sudden cardiac death. *The heart, arteries and veins* (1998), 1081–1112.

[53] EREM, B., ORELLANA, R. M., HYDE, D. E., PETERS, J. M., DUFFY, F. H., STOVICEK, P., WARFIELD, S. K., MACLEOD, R. S., TADMOR, G., AND BROOKS, D. H. Extensions to a manifold learning framework for time-series analysis on dynamic manifolds in bioelectric signals. *Physical Review E 93*, 4 (2016), 042218.

[54] EREM, B., STOVÍCEK, P., AND BROOKS, D. H. Manifold learning for analysis of low-order nonlinear dynamics in high-dimensional electrocardiographic signals. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)* (2012), IEEE, pp. 844–847.

[55] FINOCCHIARO, G., MERLO, M., SHEIKH, N., DE ANGELIS, G., PAPADAKIS, M., OLIVOTTO, I., RAPEZZI, C., CARR-WHITE, G., SHARMA, S., MESTRONI, L., ET AL. The electrocardiogram in the diagnosis and management of patients with dilated cardiomyopathy. *European journal of heart failure 22*, 7 (2020), 1097–1107.

[56] FRANCIOSI, S., PERRY, F. K., ROSTON, T. M., ARMSTRONG, K. R., CLAYDON, V. E., AND SANATANI, S. The role of the autonomic nervous system in arrhythmias and sudden cardiac death. *Autonomic Neuroscience 205* (2017), 1–11.

[57] GARCIA, T. B. *12-lead ECG: The art of interpretation*. Jones & Bartlett Publishers, 2013.

[58] GATZOULIS, K. A., ANDRIKOPOULOS, G. K., APOSTOLOPOULOS, T., SOTIROPOULOS, E., ZERVOPOULOS, G., ANTONIOU, J., BRILI, S., AND STEFANADIS, C. I. Electrical storm is an independent predictor of adverse long-term outcome in the era of implantable defibrillator therapy. *EP Europace 7*, 2 (2005), 184–192.

[59] GEMMELL, P. M., GILLETTE, K., BALABAN, G., RAJANI, R., VIGMOND, E. J., PLANK, G., AND BISHOP, M. J. A computational investigation into rate-dependant vectorcardiogram changes due to specific fibrosis patterns in non-ischæmic dilated cardiomyopathy. *Computers in Biology and Medicine 123* (aug 2020), 103895.

[60] GHISTA, D. N., ACHARYA, U. R., AND NAGENTHIRAN, T. Frontal plane vectorcardiograms: Theory and graphics visualization of cardiac health status. *Journal of medical systems 34*, 4 (2010), 445–458.

[61] GOLDBERGER, J. J., SUBAČIUS, H., PATEL, T., CUNNANE, R., AND KADISH, A. H. Sudden cardiac death risk stratification in patients with nonischemic

dilated cardiomyopathy. *Journal of the American College of Cardiology 63*, 18 (may 2014), 1879–1889.

[62] GOOD, W. W., EREM, B., ZENGER, B., COLL-FONT, J., BERGQUIST, J. A., BROOKS, D. H., AND MACLEOD, R. S. Characterizing the transient electrocardiographic signature of ischemic stress using Laplacian Eigenmaps for dimensionality reduction. *Computers in Biology and Medicine 127* (2020), 104059.

[63] GOOD, W. W., EREM, B., ZENGER, B., COLL-FONT, J., BROOKS, D. H., AND MACLEOD, R. S. Temporal performance of laplacian eigenmaps and 3D conduction velocity in detecting ischemic stress. *Journal of electrocardiology 51*, 6 (2018), S116–S120.

[64] GOOVAERTS, G., PADHY, S., VANDENBERK, B., VARON, C., WILLEMS, R., AND VAN HUFFEL, S. A machine-learning approach for detection and quantification of qrs fragmentation. *IEEE journal of biomedical and health informatics 23*, 5 (2018), 1980–1989.

[65] GOOVAERTS, G., PADHY, S., VANDENBERK, B., VARON, C., WILLEMS, R., AND VAN HUFFEL, S. Risk assessment of all-cause mortality in icd patients using a novel qrs fragmentation score. In *2018 Computing in Cardiology Conference (CinC)* (2018), vol. 45, IEEE, pp. 1–4.

[66] GOOVAERTS, G., VANDENBERK, B., VARON, C., WILLEMS, R., AND VAN HUFFEL, S. Phase-rectified signal averaging for automatic detection of QRS fragmentation. In *2016 Computing in Cardiology Conference (CinC)* (2016), IEEE, pp. 637–640.

[67] GUI, J., SUN, Z., JI, S., TAO, D., AND TAN, T. Feature selection based on structured sparsity: A comprehensive study. *IEEE transactions on neural networks and learning systems 28*, 7 (2016), 1490–1507.

[68] GUO, Y., ZHANG, Z., CAO, J., GONG, T., AND YANG, W. An optimized variational mode decomposition for extracting weak feature of viscoelastic sandwich cylindrical structures. *Measurement Science and Technology 29*, 3 (2018), 035006.

[69] GUYON, I., AND ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research 3*, Mar (2003), 1157–1182.

[70] HAIGNEY, M. C., ZAREBA, W., GENTLESK, P. J., GOLDSTEIN, R. E., ILLOVSKY, M., MCNITT, S., ANDREWS, M. L., AND MOSS, A. J. QT interval variability and spontaneous ventricular tachycardia or fibrillation in the Multicenter Automatic Defibrillator Implantation Trial (MADIT) II patients. *Journal of the American College of Cardiology 44*, 7 (2004), 1481–1487.

[71] HAN, B., TREW, M. L., AND ZGIERSKI-JOHNSTON, C. M. Cardiac Conduction Velocity, Remodeling and Arrhythmogenesis. *Cells 10*, 11 (oct 2021), 2923.

[72] HANNUN, A. Y., RAJPURKAR, P., HAGHPANAHI, M., TISON, G. H., BOURN, C., TURAKHIA, M. P., AND NG, A. Y. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine 25*, 1 (2019), 65–69.

[73] HASAN, M. A., AND ABBOTT, D. A review of beat-to-beat vectorcardiographic (VCG) parameters for analyzing repolarization variability in ECG signals. *Biomedical Engineering/Biomedizinische Technik 61*, 1 (2016), 3–17.

[74] HAUKILAHTI, M. A. E., ERANTI, A., KENTTÄ, T., AND HUIKURI, H. V. QRS Fragmentation Patterns Representing Myocardial Scar Need to Be Separated from Benign Normal Variants: Hypotheses and Proposal for Morphology based Classification. *Frontiers in Physiology 7* (2016), 653.

[75] HE, X., CAI, D., AND NIYOGI, P. Laplacian score for feature selection. In *Advances in neural information processing systems* (2006), pp. 507–514.

[76] HELDEWEG, M. L. A., LIU, N., KOH, Z. X., FOOK-CHONG, S., LYE, W. K., HARMS, M., AND ONG, M. E. H. A novel cardiovascular risk stratification model incorporating ecg and heart rate variability for patients presenting to the emergency department with chest pain. *Critical Care 20*, 1 (2016), 1–9.

[77] HNATKOVA, K., ANDRŠOVÁ, I., TOMAN, O., SMETANA, P., HUSTER, K. M., ŠIŠÁKOVÁ, M., BARTHEL, P., NOVOTNÝ, T., SCHMIDT, G., AND MALIK, M. Spatial distribution of physiologic 12-lead QRS complex. *Scientific reports 11*, 1 (2021), 1–20.

[78] HOHNLOSER, S. H., AL-KHALIDI, H. R., PRATT, C. M., BRUM, J. M., TATLA, D. S., TCHOU, P., AND DORIAN, P. Electrical storm in patients with an implantable defibrillator: incidence, features, and preventive therapy: insights from a randomized trial. *European heart journal 27*, 24 (2006), 3027–3032.

[79] HOLMSTRÖM, L., HAUKILAHTI, A., VÄHÄTALO, J., KENTTÄ, T., APPEL, H., KIVINIEMI, A., PAKANEN, L., HUIKURI, H. V., MYERBURG, R. J., AND JUNTTILA, J. Electrocardiographic associations with myocardial fibrosis among sudden cardiac death victims. *Heart 106*, 13 (2020), 1001–1006.

[80] HOU, C., NIE, F., LI, X., YI, D., AND WU, Y. Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *IEEE Transactions on Cybernetics 44*, 6 (2013), 793–804.

[81] HWANG, S.-G. Cauchy's interlace theorem for eigenvalues of hermitian matrices. *The American Mathematical Monthly 111*, 2 (2004), 157–159.

[82] INSTITUTE OF MEDICINE (US) AND NATIONAL ACADEMIES PRESS (US). *Cardiovascular Disability: Updating the Social Security Listings*. National Academies Press, 2010.

[83] IP, J. E., AND LERMAN, B. B. Idiopathic malignant premature ventricular contractions. *Trends in cardiovascular medicine 28*, 4 (2018), 295–302.

[84] Health informatics – Standard communication protocol – Part 91064: Computer-assisted electrocardiography. Standard, International Organization for Standardization, Geneva, CH, 2009.

[85] JACOBS, B., VILLA, A., MOEYERSONS, J., VAN HUFFEL, S., WILLEMS, R., AND VARON, C. Can Laplacian Eigenmaps be used for differentiation between healthy subjects and patients with corrected Tetralogy of Fallot? In *2020 Computing in Cardiology* (2020), IEEE, pp. 1–4.

[86] JAROS, R., MARTINEK, R., AND DANYS, L. Comparison of Different Electrocardiography with Vectorcardiography Transformations. *Sensors 19*, 14 (jul 2019), 3072.

[87] JEFFERIES, J. L., AND TOWBIN, J. A. Dilated cardiomyopathy. *The Lancet 375*, 9716 (2010), 752–762.

[88] JIANG, X., ZHANG, L., ZHAO, Q., AND ALBAYRAK, S. ECG arrhythmias recognition system based on independent component analysis feature extraction. In *TENCON 2006-2006 IEEE Region 10 Conference* (2006), IEEE, pp. 1–4.

[89] JIN, F., SUGAVANESWARAN, L., KRISHNAN, S., AND CHAUHAN, V. S. Quantification of fragmented QRS complex using intrinsic time-scale decomposition. *Biomedical Signal Processing and Control 31* (2017), 513–523.

[90] KARDYS, I., KORS, J. A., VAN DER MEER, I. M., HOFMAN, A., VAN DER KUIP, D. A., AND WITTEMAN, J. C. Spatial QRS-T angle predicts cardiac death in a general population. *European Heart Journal 24*, 14 (jul 2003), 1357–1364.

[91] KARSIKAS, M., NOPONEN, K., TULPPO, M., H, H., AND SEPPÄNEN, T. Beat-to-beat variation of three-dimensional QRS-T angle measures during exercise test. In *36th Annual Computers in Cardiology Conference (CinC)* (2009), pp. 125–128.

[92] KAWASAKI, T., HARIMOTO, K., HONDA, S., SATO, Y., YAMANO, M., MIKI, S., AND KAMITANI, T. Notched qrs for the assessment of myocardial fibrosis in hypertrophic cardiomyopathy. *Circulation Journal* (2015), CJ–14.

[93] KENNEDY, H. L., WHITLOCK, J. A., SPRAGUE, M. K., KENNEDY, L. J., BUCKINGHAM, T. A., AND GOLDBERG, R. J. Long-term follow-up of asymptomatic healthy subjects with frequent and complex ventricular ectopy. *New England Journal of Medicine 312*, 4 (jan 1985), 193–197.

[94] KORS, J. A., KARDYS, I., VAN DER MEER, I. M., AND VAN HERPEN, G. Spatial QRS-T angle as a risk indicator of cardiac death in an elderly population. *Journal of Electrocardiology 36* (2003), 113–114.

[95] KORS, J. A., VAN HERPEN, G., SITTIG, A. C., AND VAN BEMMEL, J. H. Reconstruction of the Frank vectorcardiogram from standard electrocardiographic leads: diagnostic comparison of different methods. *European Heart Journal 11*, 12 (dec 1990), 1083–1092.

[96] KOSSMANN, C. E., BRODY, D. A., BURCH, G. E., HECHT, H. H., JOHNSTON, F. D., KAY, C., LEPESCHKIN, E., PIPBERGER, H. V., PIPBERGER, H. V., BAULE, G., ET AL. Recommendations for standardization of leads and of specifications for instruments in electrocardiography and vectorcardiography. *Circulation 35*, 3 (1967), 583–602.

[97] KRONBORG, M. B., NIELSEN, J. C., AND MORTENSEN, P. T. Electrocardiographic patterns and long-term clinical outcome in cardiac resynchronization therapy. *EP Europace 12*, 2 (11 2009), 216–222.

[98] KURL, S., MÄKIKALLIO, T. H., RAUTAHARJU, P., KIVINIEMI, V., AND LAUKKANEN, J. A. Duration of QRS complex in resting electrocardiogram is a predictor of sudden cardiac death in men. *Circulation 125*, 21 (2012), 2588–2594.

[99] LEE, H., SHIN, S.-Y., SEO, M., NAM, G.-B., AND JOO, S. Prediction of ventricular tachycardia one hour before occurrence using artificial neural networks. *Scientific reports 6*, 1 (2016), 1–7.

[100] LI, J., CHENG, K., WANG, S., MORSTATTER, F., TREVINO, R. P., TANG, J., AND LIU, H. Feature selection: A data perspective. *ACM Computing Surveys (CSUR) 50*, 6 (2017), 1–45.

[101] LI, Z., YANG, Y., LIU, J., ZHOU, X., AND LU, H. Unsupervised feature selection using nonnegative spectral analysis. In *Twenty-Sixth AAAI Conference on Artificial Intelligence* (2012).

[102] LIN, C.-C., AND YANG, C.-M. Heartbeat classification using normalized RR intervals and morphological features. *Mathematical Problems in Engineering 2014* (2014).

[103] LIN, Y.-H., LIN, L.-Y., CHEN, Y.-S., HUANG, H.-C., LEE, J.-K., HO, Y.-L., LIAO, L.-C., AND CHEN, W.-J. The association between T-Wave morphology and life-threatening ventricular tachyarrhythmias in patients with congestive heart failure. *Pacing and Clinical Electrophysiology 32*, 9 (sep 2009), 1173–1177.

[104] LIU, W., CAO, S., AND CHEN, Y. Applications of variational mode decomposition in seismic time-frequency analysis. *Geophysics 81*, 5 (2016), V365–V378.

[105] LUNDBERG, S. M., AND LEE, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems 30* (2017).

[106] LUNDBERG, S. M. E. A. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering 2*, 10 (2018), 749–760.

[107] LUNGA, D., PRASAD, S., CRAWFORD, M. M., AND ERSOY, O. Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. *IEEE Signal Processing Magazine 31*, 1 (2013), 55–66.

[108] MAHESHWARI, S., ACHARYYA, A., PUDDU, P. E., MAZOMENOS, E. B., LEEKHA, G., MAHARATNA, K., AND SCHIARITI, M. An automated algorithm for online detection of fragmented QRS and identification of its various morphologies. *Journal of the Royal Society Interface 10*, 89 (2013), 20130761.

[109] MALIK, M. Electrocardiographic smoke signals of fragmented QRS complex. *Journal of cardiovascular electrophysiology 24*, 11 (2013), 1267–1270.

[110] MALIK, M., JOHN CAMM, A., THOMAS BIGGER, J., BREITHARDT, G., CERUTTI, S., COHEN, R. J., COUMEL, P., FALLEN, E. L., KENNEDY, H. L., KLEIGER, R. E., LOMBARDI, F., MALLIANI, A., MOSS, A. J., ROTTMAN, J. N., SCHMIDT, G., SCHWARTZ, P. J., AND SINGER, D. H. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation 93*, 5 (1996), 1043–1065.

[111] MALMIVUO, J., PLONSEY, R., ET AL. *Bioelectromagnetism: principles and applications of bioelectric and biomagnetic fields*. Oxford University Press, USA, 1995.

[112] MAN, S., MAAN, A. C., SCHALIJ, M. J., AND SWENNE, C. A. Vectorcardiographic diagnostic & prognostic information derived from the 12-lead electrocardiogram: Historical review and clinical perspective. 463–475.

[113] MAN, S., RAHMATTULLA, C., MAAN, A. C., VAN DER PUTTEN, N. H., DIJK, W. A., VAN ZWET, E. W., VAN DER WALL, E. E., SCHALIJ, M. J., GORGELS, A. P., AND SWENNE, C. A. Acute coronary syndrome with a totally occluded culprit artery: Relation of the ST injury vector with ST-elevation and non-ST elevation ECGs. In *Journal of Electrocardiology* (2014), vol. 47, Churchill Livingstone Inc., pp. 183–190.

[114] MAN, S., VAN ZWET, E. W., MAAN, A. C., SCHALIJ, M. J., AND SWENNE, C. A. Individually improved vcg synthesis. In *2009 36th Annual Computers in Cardiology Conference (CinC)* (2009), IEEE, pp. 277–280.

[115] MAR, T., ZAUNSEDER, S., MARTÍNEZ, J. P., LLAMEDO, M., AND POLL, R. Optimization of ecg classification by means of feature selection. *IEEE transactions on Biomedical Engineering 58*, 8 (2011), 2168–2177.

[116] MARTÍNEZ, J. P., ALMEIDA, R., OLMOS, S., ROCHA, A. P., AND LAGUNA, P. A wavelet-based ECG delineator: evaluation on standard databases. *IEEE Transactions on biomedical engineering 51*, 4 (2004), 570–581.

[117] M.C. DE BRUYNE, A.W. HOES, J.A. KORS, A. HOFMAN, J.H. VAN BEMMEL, D. G. Prolonged QT interval predicts cardiac and all-cause mortality in the elderly: The Rotterdam Study. *European Heart Journal 20*, 4 (1999), 278–284.

[118] MCCROHON, J. A., MOON, J. C., PRASAD, S. K., MCKENNA, W. J., LORENZ, C. H., COATS, A. J., AND PENNELL, D. J. Differentiation of heart failure related to dilated cardiomyopathy and coronary artery disease using gadolinium-enhanced cardiovascular magnetic resonance. *Circulation 108*, 1 (jul 2003), 54–59.

[119] MEMBERS, A. F. 2015 esc guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: The task force for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death of the european society of cardiology (esc) endorsed by: Association for european paediatric and congenital cardiology (aepc). *Ep Europace 17*, 11 (2015), 1601–1687.

[120] MJAAVATTEN, A. Curvature of a 1D curve in a 2D or 3D space. MATLAB Central File Exchange, 2021. Retrieved December 9, 2021.

[121] MOEYERSONS, J. *Advanced Tools for Ambulatory ECG and Respiratory Analysis*. PhD thesis, KU Leuven, 2021.

[122] MOEYERSONS, J., AMONI, M., HUFFEL, S. V., WILLEMS, R., AND VARON, C. R-DECO: An open-source Matlab based graphical user interface for the detection and correction of R-peaks. *PeerJ Computer Science 2019*, 10 (oct 2019), e226.

[123] MOEYERSONS, J., MORALES, J., VILLA, A., CASTRO, I., TESTELMANS, D., BUYSE, B., VAN HOOF, C., WILLEMS, R., VAN HUFFEL, S., AND VARON, C. Supervised svm transfer learning for modality-specific artefact detection in ecg. *Sensors 21*, 2 (2021), 662.

[124] MOEYERSONS, J., SMETS, E., MORALES, J., VILLA, A., DE RAEDT, W., TESTELMANS, D., BUYSE, B., VAN HOOF, C., WILLEMS, R., VAN HUFFEL, S., ET AL. Artefact detection and quality assessment of ambulatory ECG signals. *Computer methods and programs in biomedicine 182* (2019), 105050.

[125] MONDÉJAR-GUERRA, V., NOVO, J., ROUCO, J., PENEDO, M. G., AND ORTEGA, M. Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers. *Biomedical Signal Processing and Control 47* (2019), 41–48.

[126] MOODY, G. B., AND MARK, R. G. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine 20*, 3 (2001), 45–50.

[127] MÜLLER, M. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.

[128] MUSER, D. E. A. Risk stratification of patients with apparently idiopathic premature ventricular contractions: A multicenter international CMR registry. *JACC: Clinical Electrophysiology 6*, 6 (jun 2020), 722–735.

[129] NADOUR, W., DOYLE, M., WILLIAMS, R. B., RAYARAO, G., GRANT, S. B., THOMPSON, D. V., YAMROZIK, J. A., AND BIEDERMAN, R. W. Does the presence of Q waves on the EKG accurately predict prior myocardial infarction when compared to cardiac magnetic resonance using late gadolinium enhancement? A cross-population study of noninfarct vs infarct patients. *Heart Rhythm 11*, 11 (2014), 2018–2026.

[130] NARAYANAN, A. M., AND BERTRAND, A. Analysis of miniaturization effects and channel selection strategies for EEG sensor networks with application to auditory attention detection. *IEEE Transactions on Biomedical Engineering 67*, 1 (2020), 234–244.

[131] NELWAN, S. P., KORS, J. A., AND MEIJ, S. H. Minimal lead sets for reconstruction of 12-lead electrocardiograms. *Journal of electrocardiology 33* (2000), 163–166.

[132] NG, A. Y., JORDAN, M. I., AND WEISS, Y. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems* (2002), pp. 849–856.

[133] NIE, F., ZHU, W., AND LI, X. Structured graph optimization for unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[134] NOVOSEL, D., NOLL, G., AND LÜSCHER, T. Corrected formula for the calculation of the electrical heart axis. *Croatian Medical Journal 40*, 1 (1999), 77–79.

[135] OEHLER, A., FELDMAN, T., HENRIKSON, C. A., AND TERESHCHENKO, L. G. QRS-T angle: A review. *Annals of Noninvasive Electrocardiology 19*, 6 (2014), 534–542.

[136] OZENNE, B., SUBTIL, F., AND MAUCORT-BOULCH, D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology 68*, 8 (2015), 855–859.

[137] PANG, B. J.-M., AND GREEN, M. S. Epidemiology of ventricular tachyarrhythmia. *Herzschrittmachertherapie+ Elektrophysiologie 28*, 2 (2017), 143–148.

[138] PAVRI, B. B. E. A. Prognostic value and temporal behavior of the planar QRS-T angle in patients with nonischemic cardiomyopathy. *Circulation 117*, 25 (2008), 3181–3186.

[139] PELLI, A., KENTTÄ, T. V., JUNTTILA, M. J., BERGAU, L., ZABEL, M., MALIK, M., REICHLIN, T., WILLEMS, R., VOS, M. A., HARDEN, M., ET AL. Electrocardiogram as a predictor of survival without appropriate shocks in primary prophylactic ICD patients: A retrospective multi-center study. *International journal of cardiology 309* (2020), 78–83.

[140] PLATT, J., ET AL. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers 10*, 3 (1999), 61–74.

[141] POORANACHANDRAN, V., NICOLSON, W., VALI, Z., LI, X., AND NG, G. A. Non-invasive markers for sudden cardiac death risk stratification in dilated cardiomyopathy. *Heart* (oct 2021), heartjnl–2021–319971.

[142] RAJENDRA ACHARYA, U., SURI, J. S., SPAAN, J. A., AND KRISHNAN, S. M. *Advances in cardiac signal processing*. Springer Berlin Heidelberg, 2007.

[143] REMESEIRO, B., AND BOLON-CANEDO, V. A review of feature selection methods in medical applications. *Computers in biology and medicine 112* (2019), 103375.

[144] REYNOLDS, B. E. Taxicab geometry. *Pi Mu Epsilon Journal 7*, 2 (1980), 77–88.

[145] RIJNBEEK, P. R., VAN HERPEN, G., BOTS, M. L., MAN, S., VERWEIJ, N., HOFMAN, A., HILLEGE, H., NUMANS, M. E., SWENNE, C. A., WITTEMAN, J. C., AND KORS, J. A. Normal values of the electrocardiogram for ages 16-90 years. *Journal of Electrocardiology 47*, 6 (nov 2014), 914–921.

[146] RUBULIS, A., JENSEN, J., LUNDAHL, G., TAPANAINEN, J., WECKE, L., AND BERGFELDT, L. T vector and loop characteristics in coronary artery disease and during acute ischemia. *Heart Rhythm 1*, 3 (sep 2004), 317–325.

[147] RUBULIS, A., JENSEN, S. M., NÄSLUND, U., LUNDAHL, G., JENSEN, J., AND BERGFELDT, L. Ischemia-induced repolarization response in relation to the size and location of the ischemic myocardium during short-lasting coronary occlusion in humans. *Journal of Electrocardiology 43*, 2 (mar 2010), 104–112.

[148] SAHOO, S., MOHANTY, M., BEHERA, S., AND SABUT, S. K. Ecg beat classification using empirical mode decomposition and mixture of features. *Journal of medical engineering & technology 41*, 8 (2017), 652–661.

[149] SANDERCOCK, G. R., BROMLEY, P. D., AND BRODIE, D. A. The reliability of short-term measurements of heart rate variability. 238–247.

[150] SASSI, R., CERUTTI, S., LOMBARDI, F., MALIK, M., HUIKURI, H. V., PENG, C.-K., SCHMIDT, G., YAMAMOTO, Y., REVIEWERS:, D., GORENEK, B., ET AL. Advances in heart rate variability signal analysis: joint position statement by the e-cardiology esc working group and the european heart rhythm association co-endorsed by the asia pacific heart rhythm society. *Ep Europace 17*, 9 (2015), 1341–1353.

[151] SCHALIJ, M. J., BORLEFFS, C. J. W., SCHERPTONG, R. W., MAN, S. C., VAN WELSENES, G. H., BAX, J. J., VAN ERVEN, L., AND SWENNE, C. A. Predicting ventricular arrhythmias in patients with ischemic heart disease: Clinical application of the ECG-derived QRS-T angle. *Circulation: Arrhythmia and Electrophysiology 2*, 5 (oct 2009), 548–554.

[152] SCHNEIDER, R., ET AL. Phase-rectified signal averaging for the detection of quasi-periodicities in electrocardiogram. In *IFMBE Proceedings* (2007), vol. 16, Springer Verlag, pp. 38–41.

[153] SCHREURS, C. A., ALGRA, A. M., MAN, S. C., CANNEGIETER, S. C., VAN DER WALL, E. E., SCHALIJ, M. J., KORS, J. A., AND SWENNE, C. A. The spatial QRS-T angle in the Frank vectorcardiogram: accuracy of estimates derived from the 12-lead electrocardiogram. *Journal of Electrocardiology 43*, 4 (jul 2010), 294–301.

[154] SCHULTHEISS, H.-P., FAIRWEATHER, D., CAFORIO, A. L., ESCHER, F., HERSHBERGER, R. E., LIPSHULTZ, S. E., LIU, P. P., MATSUMORI, A.,

MAZZANTI, A., MCMURRAY, J., ET AL. Dilated cardiomyopathy. *Nature reviews Disease primers 5*, 1 (2019), 1–19.

[155] SEDAGHAT, G., GHAFOORI, E., WAKS, J. W., KABIR, M. M., SHVILKIN, A., JOSEPHSON, M. E., AND TERESHCHENKO, L. G. Quantitative assessment of vectorcardiographic loop morphology. *Journal of Electrocardiology 49*, 2 (mar 2016), 154–163.

[156] SEKO, Y., KATO, T., YAMAJI, Y., HARUNA, Y., NAKANE, E., HARUNA, T., AND INOKO, M. Clinical impact of left and right axis deviations with narrow QRS complex on 3-year outcomes in a hospital-based population in Japan. *Scientific reports 11*, 1 (2021), 1–7.

[157] SHA, J., ZHANG, S., TANG, M., CHEN, K., ZHAO, X., AND WANG, F. Fragmented QRS is associated with all-cause mortality and ventricular arrhythmias in patient with idiopathic dilated cardiomyopathy. *Annals of Noninvasive Electrocardiology 16*, 3 (2011), 270–275.

[158] SILVERTHORN, D., O. W., AND GARRISON, C. *Human Physiology: An Integrated Approach*. Pearson Higher Ed, 2004.

[159] SIONTIS, K. C., NOSEWORTHY, P. A., ATTIA, Z. I., AND FRIEDMAN, P. A. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology 18*, 7 (2021), 465–478.

[160] SORIA, M. L., AND MARTINEZ, J. An ECG classification model based on multilead wavelet transform features. In *2007 Computers in Cardiology* (2007), IEEE, pp. 105–108.

[161] STICHERLING, C., ARENDACKA, B., SVENDSEN, J. H., WIJERS, S., FRIEDE, T., STOCKINGER, J., DOMMASCH, M., MERKELY, B., WILLEMS, R., LUBINSKI, A., ET AL. Sex differences in outcomes of primary prevention implantable cardioverter-defibrillator therapy: combined registry data from eleven european countries. *Ep Europace 20*, 6 (2018), 963–970.

[162] STRAUSS, D. G., OLSON, C. W., WU, K. C., HEIBERG, E., PERSSON, E., SELVESTER, R. H., PAHLM, O., AND ARHEDEN, H. Vectorcardiogram synthesized from the 12-lead electrocardiogram to image ischemia. *Journal of Electrocardiology 42*, 2 (mar 2009), 190–197.

[163] SUR, S., HAN, L., AND TERESHCHENKO, L. G. Comparison of sum absolute QRST integral, and temporal variability in depolarization and repolarization, measured by dynamic vectorcardiography approach, in healthy men and women. *PloS one 8*, 2 (2013), e57175.

[164] SUYKENS, J. A., VAN GESTEL, T., DE BRABANTER, J., DE MOOR, B., AND VANDEWALLE, J. P. *Least squares support vector machines*. World scientific, 2002.

[165] SZURLEY, J., BERTRAND, A., AND MOONEN, M. Efficient computation of microphone utility in a wireless acoustic sensor network with multi-channel Wiener filter based noise reduction. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2012), IEEE, pp. 2657–2660.

[166] SZURLEY, J., BERTRAND, A., RUCKEBUSCH, P., MOERMAN, I., AND MOONEN, M. Greedy distributed node selection for node-specific signal estimation in wireless sensor networks. *Signal Processing 94* (2014), 57–73.

[167] TAKIGAWA, M., NODA, T., KURITA, T., AIHARA, N., YAMADA, Y., OKAMURA, H., SATOMI, K., SUYAMA, K., SHIMIZU, W., AND KAMAKURA, S. Predictors of electrical storm in patients with idiopathic dilated cardiomyopathy–how to stratify the risk of electrical storm. *Circulation Journal 74*, 9 (2010), 1822–1829.

[168] TANG, C., LIU, X., LI, M., WANG, P., CHEN, J., WANG, L., AND LI, W. Robust unsupervised feature selection via dual self-representation and manifold regularization. *Knowledge-Based Systems 145* (2018), 109–120.

[169] TERESHCHENKO, L. G., HAN, L., CHENG, A., MARINE, J. E., SPRAGG, D. D., SINHA, S., DALAL, D., CALKINS, H., TOMASELLI, G. F., AND BERGER, R. D. Beat-to-beat three-dimensional ECG variability predicts ventricular arrhythmia in ICD recipients. *Heart Rhythm 7*, 11 (2010), 1606–1613.

[170] TERESHCHENKO, L. G., WAKS, J. W., KABIR, M., GHAFOORI, E., SHVILKIN, A., AND JOSEPHSON, M. E. Analysis of speed, curvature, planarity and frequency characteristics of heart vector movement to evaluate the electrophysiological substrate associated with ventricular tachycardia. *Computers in Biology and Medicine 65* (oct 2015), 150–160.

[171] THARWAT, A., HASSANIEN, A. E., AND ELNAGHI, B. E. A BA-based algorithm for parameter optimization of support vector machine. *Pattern Recognition Letters 93* (2017), 13–22.

[172] TORIGOE, K., TAMURA, A., KAWANO, Y., SHINOZAKI, K., KOTOKU, M., AND KADOTA, J. The number of leads with fragmented QRS is independently associated with cardiac death or hospitalization for heart failure in patients with prior myocardial infarction. *Journal of cardiology 59*, 1 (2012), 36–41.

[173] TSIRONIS, S., SOZIO, M., VAZIRGIANNIS, M., AND POLTECHNIQUE, L. Accurate spectral clustering for community detection in map-reduce. In *Advances in Neural Information Processing Systems (NIPS) Workshops* (2013), Citeseer, p. 8.

[174] VANDENBERK, B. *Non-Invasive Risk Stratification of Ventricular Arrhythmia and Sudden Cardiac Death*. PhD thesis, 2017.

[175] VANDENBERK, B., ROBYNS, T., GOOVAERTS, G., CLAEYS, M., HELSEN, F., VAN SOEST, S., GARWEG, C., ECTOR, J., VAN HUFFEL, S., AND WILLEMS, R. Inter-and intra-observer variability of visual fragmented qrs scoring in ischemic and non-ischemic cardiomyopathy. *Journal of electrocardiology 51*, 3 (2018), 549–554.

[176] VANDENBERK, B., ROBYNS, T., GOOVAERTS, G., VAN SOEST, S., FLORÉ, V., GARWEG, C., VAN HUFFEL, S., ECTOR, J., AND WILLEMS, R. Inferior and anterior QRS fragmentation have different prognostic value in patients who received an implantable defibrillator in primary prevention of sudden cardiac death. *International Journal of Cardiology 243* (2017), 223–228.

[177] VARON, C. *Mining the ECG: Algorithms and applications*. PhD thesis, KU Leuven, 2015.

[178] VARON, C., ALZATE, C., AND SUYKENS, J. A. Noise level estimation for model selection in kernel PCA denoising. *IEEE transactions on neural networks and learning systems 26*, 11 (2015), 2650–2663.

[179] VERLEYSEN, M., AND FRANÇOIS, D. The curse of dimensionality in data mining and time series prediction. In *International Work-Conference on Artificial Neural Networks* (2005), Springer, pp. 758–770.

[180] VILLA, A., DEVIAENE, M., WILLEMS, R., VAN HUFFEL, S., AND VARON, C. Are we training our heartbeat classification algorithms properly? In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (2019), Institute of Electrical and Electronics Engineers Inc., pp. 6363–6366.

[181] VILLA, A., INGELAERE, S., JACOBS, B., VANDENBERK, B., VAN HUFFEL, S., WILLEMS, R., AND VARON, C. A unified framework for multi-lead ECG characterization using Laplacian Eigenmaps. *Under review in Journal of Biomedical and Health Informatics*.

[182] VILLA, A., INGELAERE, S., VAN HUFFEL, S., WILLEMS, R., AND VARON, C. Fragmented QRS dynamics towards electrical storm in ICD Patients. In *Computing in Cardiology* (2020), vol. 2020-Septe, IEEE Computer Society.

[183] VILLA, A., NARAYANAN, A. M., VAN HUFFEL, S., BERTRAND, A., AND VARON, C. Utility metric for unsupervised feature selection. *PeerJ Computer Science 7* (2021), 1–26.

[184] VILLA, A., PADHY, S., WILLEMS, R., VAN HUFFEL, S., AND VARON, C. Variational mode decomposition features for heartbeat classification. In *Computing in Cardiology* (2018), vol. 2018-Septe, IEEE Computer Society.

[185] VILLA, A., VANDENBERK, B., KENTTÄ, T., INGELAERE, S., HUIKUR, H. V., ZABEL, M., FRIEDE, T., STICHERLING, C., TUINENBURG, A., MALIK, M., VAN HUFFEL, S., WILLEMS, R., AND VARON, C. A machine learning algorithm for electrocardiographic fQRS quantification validated on multi-center data. *Under review in Scientific Reports*.

[186] VON LUXBURG, U. A tutorial on spectral clustering. *Statistics and computing 17*, 4 (2007), 395–416.

[187] VOSS, A. E. A. The application of methods of non-linear dynamics for the improved and predictive recognition of patients threatened by sudden cardiac death. *Cardiovascular research 31*, 3 (1996), 419–433.

[188] VOZDA, M., AND CERNY, M. Methods for derivation of orthogonal leads from 12-lead electrocardiogram: A review. *Biomedical Signal Processing and Control 19* (may 2015), 23–34.

[189] VULLINGS, R., PETERS, C., MOSSAVAT, I., OEI, S., AND BERGMANS, J. Bayesian approach to patient-tailored vectorcardiography. *IEEE Transactions on Biomedical Engineering 57*, 3 (2010), 586–595. cited By 13.

[190] WAGNER, P., STRODTHOFF, N., BOUSSELJOT, R.-D., KREISELER, D., LUNZE, F. I., SAMEK, W., AND SCHAEFFTER, T. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data 7*, 1 (2020), 1–15.

[191] WALTON, W. Feret's statistical diameter as a measure of particle size. *Nature 162*, 4113 (1948), 329–330.

[192] WOLD, S., ESBENSEN, K., AND GELADI, P. Principal component analysis. *Chemometrics and intelligent laboratory systems 2*, 1-3 (1987), 37–52.

[193] WONG, C. X., BROWN, A., LAU, D. H., CHUGH, S. S., ALBERT, C. M., KALMAN, J. M., AND SANDERS, P. Epidemiology of sudden cardiac death: global and regional perspectives. *Heart, Lung and Circulation 28*, 1 (2019), 6–14.

[194] YAMAZAKI, T., FROELICHER, V. F., MYERS, J., CHUN, S., AND WANG, P. Spatial QRS-T angle predicts cardiac death in a clinical population. *Heart Rhythm 2*, 1 (jan 2005), 73–78.

[195] YAN, S., XU, D., ZHANG, B., ZHANG, H.-J., YANG, Q., AND LIN, S. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence 29*, 1 (2006), 40–51.

[196] YANG, Y., SHEN, H. T., MA, Z., HUANG, Z., AND ZHOU, X. L2,1-norm regularized discriminative feature selection for unsupervised. In *Twenty-Second International Joint Conference on Artificial Intelligence* (2011).

[197] YE, C., KUMAR, B. V., AND COIMBRA, M. T. Heartbeat classification using morphological and dynamic features of ECG signals. *IEEE Transactions on Biomedical Engineering 59*, 10 (2012), 2930–2941.

[198] YE, C., KUMAR, B. V., AND COIMBRA, M. T. Heartbeat classification using morphological and dynamic features of ecg signals. *IEEE Transactions on Biomedical Engineering 59*, 10 (2012), 2930–2941.

[199] YIN, S., AND YIN, J. Tuning kernel parameters for SVM based on expected square distance ratio. *Information Sciences 370* (2016), 92–102.

[200] YOU, T., LUO, C., ZHANG, K., AND ZHANG, H. Electrophysiological mechanisms underlying t-wave alternans and their role in arrhythmogenesis. *Frontiers in physiology 12* (2021), 235.

[201] YOUNG, W. J., VAN DUIJVENBODEN, S., RAMÍREZ, J., JONES, A., TINKER, A., MUNROE, P. B., LAMBIASE, P. D., AND ORINI, M. A method to minimise the impact of ECG marker inaccuracies on the spatial QRS-T angle: evaluation on 1512 manually annotated ECGs. *Biomedical signal processing and control 64* (2021), 102305.

[202] ZABEL, M., ACAR, B., KLINGENHEBEN, T., FRANZ, M. R., HOHNLOSER, S. H., AND MALIK, M. Analysis of 12-lead T-wave morphology for risk stratification after myocardial infarction. *Circulation 102*, 11 (2000), 1252–1257.

[203] ZABEL, M., WILLEMS, R., LUBINSKI, A., BAUER, A., BRUGADA, J., CONEN, D., FLEVARI, P., HASENFUSS, G., SVETLOSAK, M., HUIKURI, H. V., ET AL. Clinical effectiveness of primary prevention implantable cardioverter-defibrillators: results of the EU-CERT-ICD controlled multicentre cohort study. *European heart journal 41*, 36 (2020), 3437–3447.

[204] ZHANG, D., HE, F., HAN, S., ZOU, L., WU, Y., AND CHEN, Y. An efficient approach to directly compute the exact Hausdorff distance for 3D point sets. *Integrated Computer-Aided Engineering 24*, 3 (2017), 261–277.

[205] ZHANG, K., ALEEXENKO, V., AND JEEVARATNAM, K. Computational approaches for detection of cardiac rhythm abnormalities: Are we there yet? *Journal of electrocardiology 59* (2020), 28–34.

[206] ZHANG, Q., MANRIQUEZ, A. I., MÉDIGUE, C., PAPELIER, Y., AND SORINE, M. An algorithm for robust and efficient location of T-wave ends in electrocardiograms. *IEEE Transactions on Biomedical Engineering 53*, 12 (dec 2006), 2544–2552.

[207] ZHANG, R., NIE, F., WANG, Y., AND LI, X. Unsupervised feature selection via adaptive multimeasure fusion. *IEEE transactions on neural networks and learning systems 30*, 9 (2019), 2886–2892.

[208] ZHANG, Z., DONG, J., LUO, X., CHOI, K.-S., AND WU, X. Heartbeat classification using disease-specific feature selection. *Computers in biology and medicine 46* (2014), 79–89.

[209] ZHAO, L., HU, Q., AND WANG, W. Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso. *IEEE Transactions on Multimedia 17*, 11 (2015), 1936–1948.

[210] ZHAO, Z., AND LIU, H. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning* (2007), pp. 1151–1157.

[211] ZHU, X., LI, X., ZHANG, S., JU, C., AND WU, X. Robust joint graph sparse coding for unsupervised spectral feature selection. *IEEE transactions on neural networks and learning systems 28*, 6 (2016), 1263–1275.

Curriculum Vitae

Amalia Villa was born in Madrid, Spain, on September 10, 1992. In 2015, she received her bachelor degree in Telecommunications Engineering from Universidad Politécnica de Madrid, Spain. In September of the same year, she moved to Leuven, Belgium, to study the Master in Biomedical Engineering at KU Leuven, receiving her degree with cum laude in 2017.
She has been working since September 2017 to obtain the PhD degree at the STADIUS Center of Dynamical Systems, Signal Processing and Data Analytics (STADIUS) at the Department of Electrical Engineering (ESAT) of KU Leuven under the supervision of prof. dr. ir. Sabine Van Huffel, prof. dr. Rik Willems and prof. dr. ir. Carolina Varon.

List of publications

Articles in international journals

1. **VILLA, A.**, INGELAERE, S., JACOBS, B., VANDENBERK, B., VAN HUFFEL, S., WILLEMS, R., AND VARON, C. A unified framework for multi-lead ECG characterization using Laplacian Eigenmaps. Under review in *IEEE Journal of Biomedical and Health Informatics*.

2. **VILLA, A.**, VANDENBERK, B., KENTTÄ, T., INGELAERE, S., HUIKUR,H.V., ZABEL, M., FRIEDE, T., STICHERLING, C., TUINENBURG, A. MALIK, M., VAN HUFFEL, S., WILLEMS, R., AND VARON, C. A machine learning algorithm for electrocardiographic fQRS quantification validated on multi-center data. Under review in *Scientific Reports*.

3. MOEYERSONS, J., MORALES, J., **VILLA, A.**, CASTRO, I., TESTELMANS, D., BUYSE, B., VAN HOOF, C., WILLEMS, R., VAN HUFFEL, S. AND VARON, C. Supervised SVM Transfer Learning for Modality-Specific Artefact Detection in ECG. *Sensors 21, no. 2 (2021): 662*. https://doi.org/10.3390/s21020662.

4. **VILLA, A.**, NARAYANAN, A. M., VAN HUFFEL, S., BERTRAND, A., VARON, C. Utility metric for unsupervised feature selection. *PeerJ Computer Science* 2019. 7:e477

5. MOEYERSONS, J., SMETS, E., MORALES, J., **VILLA, A.**, DE RAEDT, W., TESTELMANS, D., BUYSE, B., VAN HOOF, C., WILLEMS, R., VAN HUFFEL, S. AND VARON, C. Artefact Detection and Quality Assessment of Ambulatory ECG Signals. *Computer Methods and Programs in Biomedicine 182, p.105050 (2019)*. https://doi.org/10.1016/j.cmpb.2019.105050.

Papers in proceedings of international conferences

1. JACOBS, B., **VILLA, A.**, MOEYERSONS, J., VAN HUFFEL, S., WILLEMS, R., & VARON, C. (2020, September). Can Laplacian Eigenmaps be used for differentiation between healthy subjects and patients with corrected Tetralogy of Fallot?. *Computing in Cardiology 47*, Rimini, Italy. (Sep 13-16, 2020).

2. **VILLA, A.**, INGELAERE, S., VAN HUFFEL, S., WILLEMS, R., & VARON, C. (2020, September). Fragmented QRS Dynamics Towards Electrical Storm in ICD Patients. *Computing in Cardiology 47*, Rimini, Italy. (Sep 13-16, 2020).

3. CASTRO, I., VARON, C., MOEYERSONS, J., **VILLA, A.**, MORALES, J., DEVIAENE, M., TORFS, T., VAN HUFFEL, S., PUERS, R., & VAN HOOF, C. Data quality assessment of capacitively-coupled ecg signals. *Computing in Cardiology 46*, Singapore, Singapore. (Sep 8-11, 2019).

4. **VILLA, A.**, DEVIAENE, M., WILLEMS, R., VAN HUFFEL, S., & VARON, C. (2019, July). Are we training our heartbeat classification algorithms properly?. In *IEEE Engineering in Medicine and Biology Society 41*. Berlin, Germany. (Jul 23-27, 2019).

5. **VILLA, A.**, PADHY, S., WILLEMS, R., VAN HUFFEL, S., & VARON, C. (2018, September). Variational mode decomposition features for heartbeat classification. *Computing in Cardiology 45*, Maastricht, The Netherlands. (Sep 23-26, 2018).

Abstracts in proceedings of (inter)national conferences

1. **VILLA, A.**, INGELAERE, S., VAN HUFFEL, S., WILLEMS, R. & VARON, C. (2020, September). Fragmented QRS Dynamics Towards Electrical Storm in ICD Patients. *Dutch Biomedical Engineering Conference (BME2021)*. Online event (Jan 2021).

2. **VILLA, A.**, NARAYANAN, A. M., VAN HUFFEL, S., BERTRAND, A. & VARON, C. Spectral Utility Feature Selector for Unsupervised Learning. *Belgian Day on Biomedical Engineering 18*. Brussels, Belgium (Nov 29, 2019).

3. **VILLA, A.**, PADHY, S., WILLEMS, R., VAN HUFFEL, S. & VARON, C. Variational Mode Decomposition features for heartbeat classification. *Dutch Biomedical Engineering Conference (BME2019)* Egmond aan Zee, The Netherlands. (Jan 24-25, 2019).

4. **VILLA, A.**, VANDENBERK, B., ROJAS HERNÁNDEZ, L., SUÁREZ LÉON, A. A., VÁZQUEZ SERISEDOS, C. R., WILLEMS, R. & VAN HUFFEL, S. A tool for automatic heartbeat classification in ambulatory *Belgian Day on Biomedical Engineering 16*. Brussels, Belgium (Dec 1, 2017).



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS
Kasteelpark Arenberg 10 box 2446
3001 Leuven