In Search of an Optimal Subset of Electrocardiogram Features to Augment the Diagnosis of Acute Coronary Syndrome at the Emergency Department

by

Zeineb Bouzid

B.S. in Electrical Engineering, INSA Lyon, 2019

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

Zeineb Bouzid

It was defended on

July 14, 2020

and approved by

Ervin Sejdić, Ph.D, Associate Professor

Department of Electrical and Computer Engineering

Murat Akcakaya, Ph.D, Assistant Professor

Department of Electrical and Computer Engineering

Ahmed Dallal, Ph.D., Assistant Professor

Department of Electrical and Computer Engineering

Salah Al-Zaiti, Ph.D & RN, Assistant Professor

School of Nursing, Department of Acute & Tertiary Care

Thesis Advisor: Ervin Sejdić, Ph.D, Associate Professor

Department of Electrical and Computer Engineering

In Search of an Optimal Subset of Electrocardiogram Features to Augment the Diagnosis of Acute Coronary Syndrome at the Emergency Department

Zeineb Bouzid, M.S.

University of Pittsburgh, 2020

The electrophysiology of acute myocardial ischemia is well understood; yet clinical practice primarily relies on classical ST amplitude measures. This translates into poor diagnostic sensitivity for identifying acute coronary syndrome (ACS). Machine learning could help identify an optimal subset of features to augment clinicians' decision during patient evaluation. We sought to compare the accuracy of supervised classifiers using electrocardiogram (ECG) feature subsets selected based on data-driven techniques or domain-specific knowledge.

This was an observational study of two prospective cohorts of consecutive patients evaluated at the emergency department for suspected ACS (Cohort 1: n=745, age 59±17, 42% Female; Cohort 2: n=499, age 59±16, 49% Female). A total of 554 temporal-spatial waveform features were extracted from baseline 12-lead ECGs using manufacturer-specific software. We used multiple algorithms to identify a subset of 229 data-driven features. Additionally, we selected a subset of 65 physiology-driven features that are mechanistically linked to myocardial ischemia. Using these two subsets of features, we evaluated logistic regression (LR) and artificial neural network (ANN) classifiers using 10-fold cross-validation on cohort 1 with independent testing on cohort 2. Our results show that classifiers with data-driven features were superior during model training (Area under the ROC curve: 0.81±0.06 vs 0.76±0.09 for LR, and 0.85±0.07 vs. 0.80±0.05 for ANN), but they generalized poorly to testing data (Area under the ROC curve: 0.68 vs 0.76 for LR, and 0.72 vs. 0.77 for ANN). In addition to classical ST and T wave amplitudes, the following features were found to be important in ACS classification: T peak–Tend interval; QRS and T axes with corresponding angles; T loop morphology, and principal component analysis ratio of ECG waveforms.

This abstract is taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

In this study, we identified a subset of novel ECG features that would improve ACS detection. These features guided by domain-specific knowledge yielded stable LR classifiers highly adaptable to be implemented in clinical decision support tools.

Keywords: machine learning, dimensionality reduction, acute coronary syndrome, electrocardiogram, ischemia.

This abstract is taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

Table of Contents

List of Tables

List of Figures

Preface

I would like to thank my supervisor Dr. Ervin Sedjić for his encouragement, full support and patient guidance during my Master studies at the University of Pittsburgh. I wish also to acknowledge the valuable help provided by our collaborators Stephanie Frisch, Ziad Faramand, and Dr. Salah Al-Zaiti, they assisted me with their clinical expertise and kind supervision to successfully complete this research project. My special thanks are extended to all my family, friends and instructors in Tunisia, France and the USA for the education and moral support I received. I specifically would like to express my deep gratitude to my parents and my partner for their unconditional love. I would like to express my very great appreciation to my undergraduate engineering school INSA Lyon for the opportunities it gave me, ending with my selection to continue my studies at the Swanson School of Engineering in the framework of a double degree agreement. I am particularly grateful for the assistance given by Pr. Claudine Gehin et Pr. Mickael Lallart who believed in me all the way.

1.0 Introduction

1.1 Acute Coronary Syndrome and Machine Learning[1]

The electrocardiogram (ECG) reflects mechanical phenomena electrically induced in the heart and is, for instance, the only tool to timely diagnose ST-segment elevation myocardial infarction for patients presented in the emergency department [2]. Clinicians are trained to formulate a diagnosis on the basis of established expert definitions of cardiac concepts and practice guidelines standardizing the usage of technologies deployed for diagnosis purposes. However, criteria defining cardiac diseases are evolving as technological breakthroughs are happening and research is advancing to improve the accuracy and sensitivity of previously existing tools, resulting in the detection of novel patterns of the human body. These exploratory achievements are acknowledged, to cite an instance, by the Joint European Society of Cardiology/American College of Cardiology committee in its attempt to redefine myocardial infarction [3]. With the benefit of hindsight, unprecedented evolutions of concepts can help better identify pathological signs in an ECG leading to a better diagnosis and a timely detection of patients suffering from a possible cardiac event and, hopefully, preventing complications to occur.

An ischemic ECG is recognized using an exploratory procedure for assessing some temporal- and spatial features of the ECG: St-elevation, left bundle branch block, Q-waves and T-waves [4]. The clinical interpretations of the changes of these features need to be done accounting for the symptoms of the patient [5]. With appropriate software, novel temporal-spatial features of the 12-lead ECG are harvested from the raw signal wave and can conceptually optimize acute coronary syndrome (ACS) detection beyond that of classical ST amplitude measurements. Machine learning techniques have proven capable of discriminating populations of ACS/non-ACS patients surpassing clinicians and approaching the performance of

[1]Portions of this part are taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

the HEART score [6]. It is then of great interest to investigate the features which had the highest impact on this data-driven decision.

Several studies have shown that, in emergency settings, the immediate recognition of ACS is a longstanding challenge [7, 8, 9]. The ECG is readily available during initial patient evaluation, and ECG markers indicative of acute myocardial ischemia have the potential to accelerate ACS diagnosis by replacing the current time-consuming biomarker-based approach [10, 11, 12]. Over the past few decades, thorough studies have explored the electrophysiological basis of acute myocardial ischemia [13], with many of them advocating the large presence of hidden signs of acute myocardial ischemia in the ECG signal [14, 15]. However, guidelines, currently effective, exclusively count on the amplitude of ST segment and T wave for ACS detection [16], rendering a diagnostic sensitivity of approximately 40% for the standard 12-lead ECG [17]. Computational algorithms open great horizons to boost the study of ECG waves by extracting numerous features from an individual 10-second 12-lead ECG. Consequently, recent advances in pattern recognition and machine learning could establish an optimal subset of features to expand clinical knowledge and enhance decision-making ability, at initial evaluation, to promptly identify ACS patients [6].

In spite of its ample employment in various clinical applications, machine learning has limitations ranging from the relatively small sample size of accessible clinical data to the requirements of replication implying the use of comparable external data sets which is usually a complicated task [18]. Accordingly, feature subset selection represents a powerful method that would substantially advance the quality of a supervised classification, comprising a better interpretability of the resulting classifier. In supplement to data-driven features selection approaches, domain-specific expertise is advised to be necessary, by a number of studies, to guide feature selection and contribute to model development during the training stage [18]. Although manual feature selection can be helpful in specifying electrophysiologically relevant characteristics of myocardial ischemia taking into account the solid knowledge background in cardiac physiology, the efficiency of this method when coupled with supervised machine learning classifiers has no sufficient proof. Indeed, manual feature selection is rather counter-intuitive to the mission of machine learning which is revealing new patterns invisible to and not captured by a human observer.

1.2 Acute Coronary Syndrome in the USA: Prevalence and Forms

Nearly 6% of the patients coming to the emergency department have a chief complaint of chest pain. Differentiating patients who have ACS versus another disease process can be challenging. The term acute coronary syndrome is an umbrella term which includes myocardial infarction, myocardial ischemia/injury and unstable angina [19].

1.2.1 Myocardial Infarction

Myocardial infarction occurs in heart cells when the body experiences prolonged ischemia (i.e., a lack of oxygen) which leads to cell death. This process of cellular death is fast because of mitochondrial abnormalities occur within 10 minutes of decrease in oxygen consumption. The reference biomarker used to detect this phenomenon is cardiac troponin I and T (abbreviated as cTnI and cTnT); a cTn level above the 99th percentile upper reference limit reveals the presence of MI injury. However, according to the evolution pattern, the latter can be classified as acute or chronic [20].

Myocardial infarction is the result of a failure in satisfying the oxygen demand. A patient going through a myocardial infarction may present with no symptoms, less frequent symptoms or more frequent ones such as chest pain and diffuse unchanged discomfort [20].

We can clinically discriminate between two classes of myocardial infarction. First, the patient's electrocardiogram can present with ST-segment elevation in two contiguous leads or 'bundle branch blocks with ischemic repolarization patterns' leading to identify it as an ST-elevation myocardial infarction [20]. Second, the absence of these signs in the ECG leads results in diagnosing the patient with a non-ST- elevation myocardial infarction [20].

1.2.2 Myocardial Ischemia/Injury

Myocardial injury of the heart muscle will have clinical evidence of acute myocardial ischemia with the detection of a rise and fall of a cardiac troponin value with at least one value above the 99th percentile upper range limit and at least one of the following characteristics: 1) symptoms of myocardial ischemia; 2) new ischemia or presumed to be

new ECG changes; 3) development of pathological Q waves; 4) imaging evidence of new loss of viable myocardium or new regional wall motion abnormality; and 5) identification of a coronary thrombus by cardiac angiogram or autopsy [16].

1.2.3 Unstable Angina

Unstable angina is a clinical diagnosis that represents when a patient has a chief complaint or potential symptoms that may be suggestive of myocardial ischemia/infarction, but lacks physiological evidence of acute myocardial death [21]. These symptoms can vary and may be, but not limited to the following: chest pain, chest discomfort, upper extremity pain, or epigastric pain lasting for greater than 20 minutes [21].

1.3 Research Objectives

In view of all the reasonings elaborated in Sections 1.1 and 1.2, we compared the effects of two feature selection approaches, domain-specific expertise versus data driven techniques, on the performance of machine learning classifiers in the specific clinical task of the electrocardiographic prediction of ACS. ECG diagnosis of ACS is a very complex task. As previously mentioned, extensive studies were conducted on acute myocardial ischemia [22, 23, 24, 25]. Still, the sensitivity of experienced clinician for ECG diagnosis of ACS is astonishingly less than 50% [6]. These findings imply that the existence of unknown signatures of myocardial ischemia hidden in the ECG signal, which suggests an important opportunity to improve diagnostics. Subsequently, we wanted to (1) compare the performance of machine learning classifiers in diagnosing ACS utilizing ECG feature subsets produced by either data-driven procedures or domain-specific know-how; and (2) whether data-driven feature selection can pinpoint ECG features suggestive of ACS and overlooked by human experts. The research to answer these questions was carried out on the available data of two prospective clinical cohorts from the EMPIRE study.

1.4 Thesis Outline

We review in Chapter 2 of the present thesis the fundamentals of an electrocardiogram and its role in predicting ACS, an outcome that will be further explained in the same chapter. Afterwards, a concise overview of the basics of machine learning is presented as an introduction to this domain.

Chapter 3 details the methodology we opted for throughout the research, putting the emphasis on the EMPIRE data collection and preprocessing alongside with the feature reduction approaches and algorithms and their efficiency in classifying patients with ACS.

The results of this study are introduced in Chapter 4 ranging from the training and validation operated on a first cohort to testing run on a separate second cohort. The outcomes of this work are given for different feature reduction approaches.

Chapter 5 consists of a prolonged discussion of the finding of this study and their entailments, investigating their potential impact on the detection of electrical biomarkers mechanistically linked to ischemia by clinicians.

Finally, we summarize our conclusions and give future insights in Chapter 6.

Portions of this thesis are taken from two forthcoming papers cited, in their current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** and **S. S. Al-Zaiti, L. Besomi, Z. Bouzid, Z. Faramand, S. O. Frisch, C. Martin-Gill,R. Gregg, S. Saba, C. Callaway, and E. Sejdi c, "Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram,"*Nature Communications*, 2020, in press.**. (Available in references [1] (submitted, under review) and [6] (submitted, under review).).

2.0 Background

2.1 Electrocardiogram

The electrocardiogram, through its leads' signals, reflects the electrical activity in the heart and has a diagnostic power that makes it very useful in the primary assessment of the patient's condition because it is non-invasive in nature and can easily be obtained. Most commonly called ECG, this tool, which can record an electrical signal of the heart muscle, was discovered by Köllicker and Müller in 1856 [26], a few decades after Galvani discovered animal electricity which served as a ground for the concept of electrophysiology [27, 28]. The evolution of the ECG on the human body was discovered by Muirhead in 1870 which has been transformed into the current date ECG by Einthoven in the early 1900's [28].

The ECG reflects the electrophysiological behavior of the cardiac cells at any given point of time, when the ECG is analyzed. All cardiac cells are positively charged on their surface due to the relative distribution of cations. The resting membrane potential of any given cardiac cell will be reverse by the process of depolarization, when the cardiac cell is stimulated. The repolarization process returns the cardiac membranes to their resting potential. These two processes are the basis for understanding and interpreting an ECG signal and collectively they are referred to as an action potential [29].

The regularity of the heart's contraction and relaxation cycle, ensured by specialized conductive cell, allows to capture the electrophysiological events at its origin using the ECG or EKG, for a better verbal distinguishability from electroencephalogram [29].

The ECG tracing of a healthy heart activity is well determined, and specific dimensions are set to define a healthy cycle, discriminating it from a misfunctioning heart. First, the isoelectric line indicates that membrane potentials are at rest and forms the baseline of an ECG [29]. The ECG intervals are indexed with letters in alphabet order announcing the beginning of an event of the cardiac cycle [29]. However, the ECG reflects only three of the sequential eight physiological events involved in the heart cycle (Figure 1) [29].

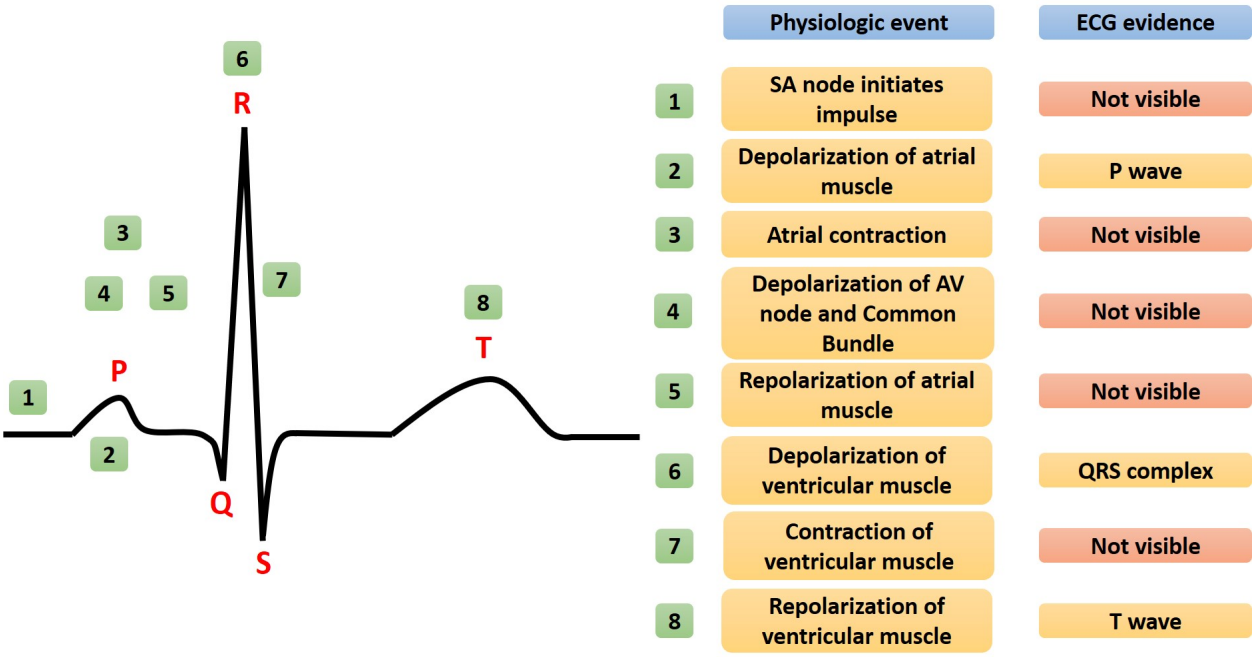

Figure 1: Cardiac cycle events. Adapted from [29].

Since the electrical activity of the heart occurs in different directions, a prior analysis led to the standardizing electrode placement to record those signals. The previously defined Einthoven's triangle, having the heart at the center of the virtual segments formed by linking the positively and negatively charged electrodes, represents nowadays the leads I, II, and III [29]. The modern ECG consists of 12 leads, still lead II is more frequently used thanks to the fact that its waves are larger than those of the other leads [29].

Leads' vectors belong to the frontal plane (I, II, III, aV_R, aV_L, and aV_F) or horizontal plane (V_1, V_2, V_3, V_4, V_5 and V_6) [30]. After fixing the electrodes on the patient's extremities at rest, Lead I, Lead II and Lead III record differences in potentials between, respectively, left arm and right arm, right arm and left leg and finally left arm and left leg, each time one of the two electrodes will be the positive pole as shown in Figure 2 [30]. The "augmented leads" aV_R, aV_L, and aV_F though have a slightly less straightforward definition since they reflect the difference in potential between one extremity's electrode and a "ground lead" which is simply the summation of the remaining two extremities leads (Figure 2) [30]. Whereas these leads are situated in the frontal plane and track the cardiac electrical activity over

its 360, V_1 through V_6 leads cover all 360 of the horizontal plane using each one distinctive electrode and the previously mentioned limbs electrodes (Figure 3) [30]. The positive pole is the particular lead electrode and the negative pole is obtained as the connection of the extremities electrodes [30].

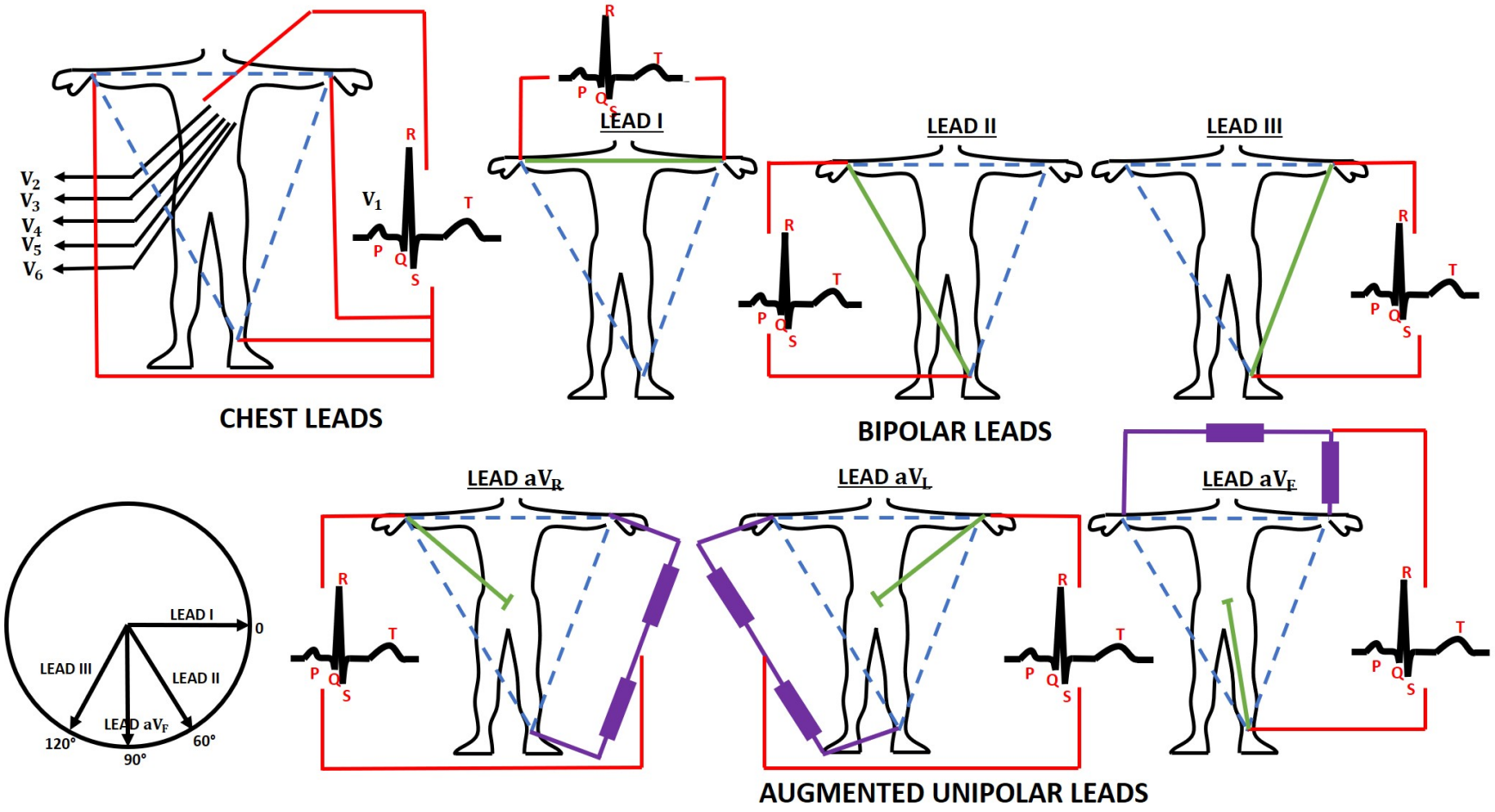

Figure 2: Details about the 12 leads of an ECG: placement and role. Adapted from [30].

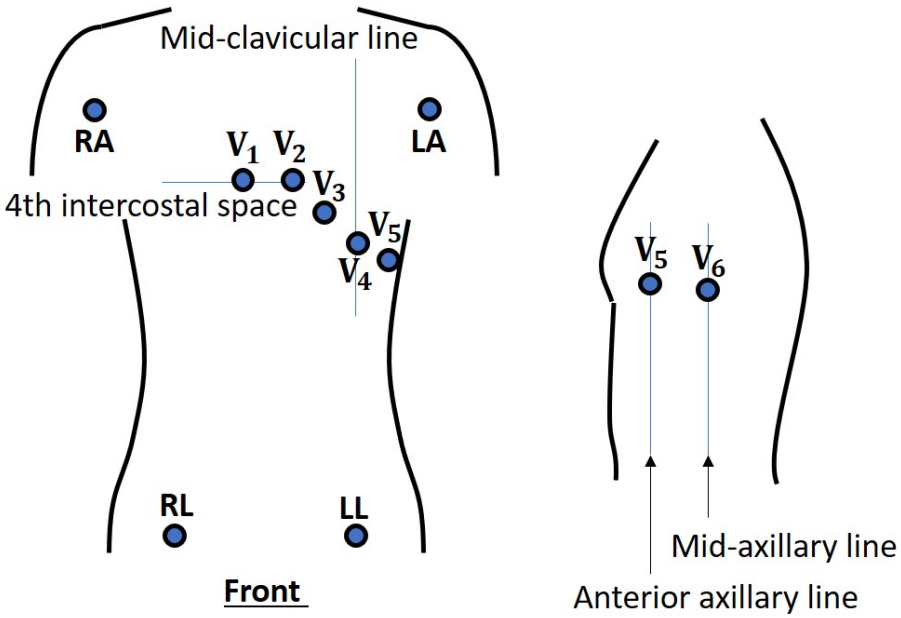

Figure 3: Specific electrodes placement to record the 12 leads. Adapted from [31].

When printed on gridded paper, the recorded signals are traced on a basis of two axis. The vertical axis indicates the voltage and direction of the electrical signals, with 1mm elementary unit length, while the horizontal axis shows the time basis as well as the sequence of cardiac cycle phases, with a 0.04 second as finest division [29].

Figure 4 is an annotated example of a beat on an ECG tracing. Letters indicate names of waves based on which intervals are defined and should respect specific dimensions in healthy individuals, these waves represent depolarization and repolarization of the heart parts [30]. Depending on heart rate, a clinician has different methods in examining an ECG and their reasoning is personal in identifying cardiac diseases such as arrythmia [29].

ECG instrumentation had to be improved to match the growth of telemedicine and e-healthcare that would lead to the necessity of the implementation of wearable ECG monitoring device [28]. The omnipresence of ECG recordings in various settings rises the premise of big data mining, and, for instance, the Web-based framework Cloudwave was developed to overcome the challenge of real-time accessibility and management of tremendous amount of

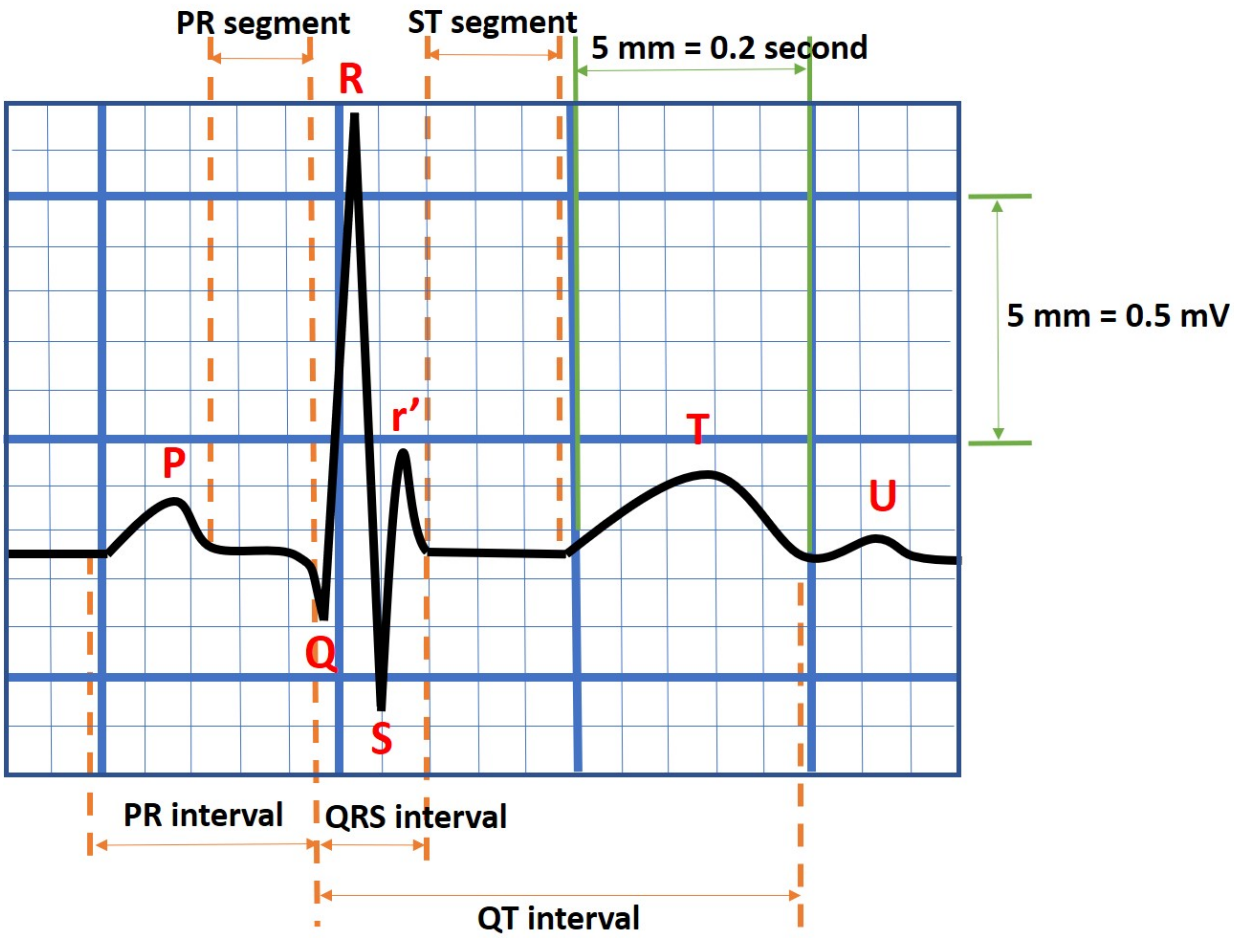

Figure 4: ECG waves and intervals. Adapted from [30].

data [28, 32]. New predictive algorithms have enhanced the diagnosis of cardiac events based on ECG, especially in the absence of a gold standard for that purpose [28]. One example is the detection of myocardial ischemia [33].

2.2 Machine Learning

2.2.1 Overview

Machine Learning as a concept emerged, with Arthur Samuel, as a trial to convey to machines the ability to learn without prior explicit programming (indicating the exact procedures to perform in specific situations) [34]. He simulates different approaches to machine learning by the use of the game of checkers briefly distinguishing between a general-purpose neural net approach, which is specifically "a randomly connected switching net", operating with a "reward-and-punishment routine" for the purpose of learning, and a special-purpose approach with an extremely organized network [34]. This potential of machines to autonomously investigate underlying patterns and make their own deductions pushed researchers to further explore this domain. In 1997, Tom Mitchell provided his rigorous definition as: "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" [35].

The concept of machine learning relies on the duality of training and testing. Following the above definition, for a process to have a learning ability, its constituent parts need to be established: the experience E (observed features used to train the model), the task T (classification, regression, clustering, etc) and the performance measure P (Area Under Curve of the Receiver Operating Characteristic, sensitivity, specificity, etc). The objective is to increase the machine's knowledge by feeding its input with the training features or observations, and then testing this learning process on a non-intersecting set of observations to whose outcomes the algorithm is completely blinded. Performance metrics, precised later in this thesis, allow the evaluation of the algorithms' efficiency.

Machine learning has expanded tremendously since it was first introduced and is now essential in advancing research in multiple fields ranging from medicine to defense and finance. Examples of applications include diagnosis and prognosis of diseases, recommender systems and autonomous vehicles.

2.2.2 Problems Types

Problems treated with machine learning lay mainly under the scope of one of the following three categories: supervised learning, unsupervised learning and reinforcement learning. However, we will also briefly inspect four additional categories, namely: semi-supervised learning, transductive inference, on-line learning and active learning.

Supervised learning occurs when the algorithm is provided with a data set with each datum representing a group of features and its corresponding actual outcome and where the program's final goal is to optimize a set of initially unknown parameters corresponding to the known features [36]. In the framework of the training stage, the goal of such algorithms is to tune a group of parameters with respect to the provided outcomes. This process is done using an optimization program that minimizes a cost function developed to reflect the correctness of the estimations for the training set compared to the actual labels. Afterwards, during the testing stage, the algorithms aim to provide the most accurate predictions for the unseen data kept separately from the training set [37].

All algorithms rely on adjusting their parameters with reference to a training set, and then testing the learnt behavior on an independent testing set. Problems defined as classification, regression or ranking problems belong to supervised learning [37]. For example, hand-written digits' recognition is a classification problem while price estimation of houses, furniture or any good or service is a regression problem.

As opposed to this "passive learning" method, active learning is a more dynamic approach by which the learning machine can interact and adapt in the process of gathering data points for training [37].

In **unsupervised learning**, called "learning without a teacher" by DeLiang Wang [36], the algorithm is only provided a set of data points as in the previous case but without their real labels. Consequently, the algorithm's role is to identify a coherent structure laying underneath observations that are not random since their provenance represents a physical process [36]. Unsupervised learning relies on two techniques as mentioned by DeLiang Wang. First, an optimization problem consisting of minimizing the entropy or maximizing the mutual information [36]. The final goal is to decrease correlation (redundancy) of the input

data [36]. Second, the Hebbian learning rule introduced by Hebb in 1949 [38] is very useful at establishing correlation since it affirms that "the connection between two neurons is strengthened if they fire at the same time (Hebb 1949)" [36].

Two major areas exploit this concept: clustering which groups data points in separate agglomerations for classification purposes and dimensionality reduction which guarantees the efficient recognition of important features to a certain application or simply helps visualize the data on the basis of certain features.

Semi-supervised learning mixes supervised and unsupervised learning features in a way that the algorithm is trained with data points that have their original real corresponding outputs as well as data points which outputs are unknown [37]. This technique is implemented in the aim of an improved performance compared to the one obtained with supervised learning alone, in cases presenting abundant observations' features and where their labels are harder to be available [37]. Similarly, transductive inference algorithms are fed with data presenting the same characteristics of the one inputted to the semi-supervised algorithm but they only test for the unlabeled data points of that same set, as an easy and realistic approach [37].

On-line learning comes with a minimization problem based on the concept of cumulating the loss obtained throughout the repeated process where it alternates training on a data point regardless of its outcome and testing on the real output of that data point [37]. Whereas it presents a similar training-testing scenario, **reinforcement learning** is characterized by the interaction the learner establishes with its environment which may lead to an action-triggered reward [37]. In the absence of a "long-term reward feedback", the exploration versus exploitation trade-off remains critical for the agent to handle as it needs to decide whether to discover the environment through new states and actions or to use the current present information to maximize its reward [37]. Besides, we can distinguish between the case of a planning problem where the environment details are familiar to the learner, and the case of a learning problem where the environment is rather anonymous, benefiting from the model of Markov decision processes [37].

2.2.3 Cross-Validation

Cross-validation enables a better understanding of the generalization effectiveness of a classifier to a separate unseen data set. The key point is eliminating the variability of the results by averaging the testing performance metric values over all the folds, but unfortunately presenting a "pessimistic bias" [39]. In principle, for a k-fold cross validation, the data set is divided into k splits, disjoint and ideally equal in size, each would be used once as the validation (testing) set and (k-1) times in the training set [39]. The split corresponding to the fold number will be kept apart for validation, and the rest of the folds will blend to form the training set [39].

Figure 5 shows the step-by-step process of the k-fold cross-validation mechanism.

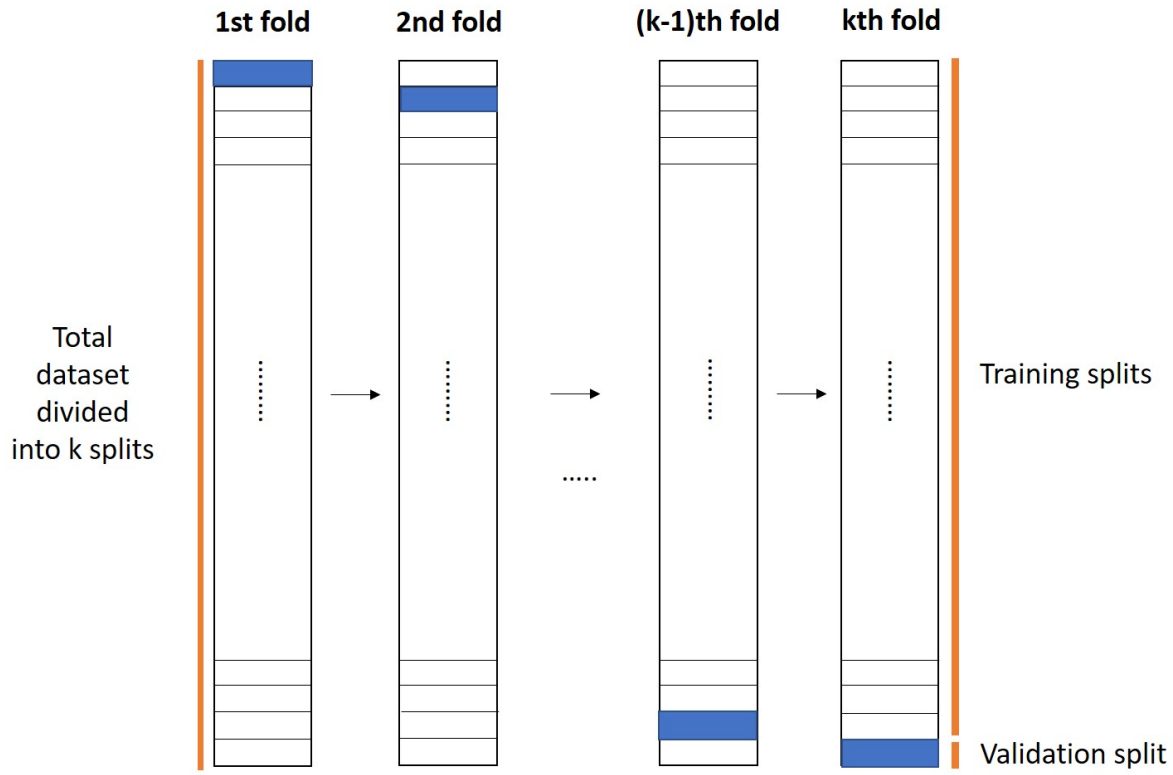

Figure 5: Schematic description of the k-fold cross-validation process

As the number of folds in a k-fold cross-validation is to be determined taking into account the total number of observations available, one way to apply this method is to consider k as equal to the total data set size, this method is called leave-one-out cross-validation [40, 41, 39]. It is recommended for a small sample size [42].

K-fold cross-validation is particularly useful for the assessment of the generalization performance of a model as well as for the choice of the most adequate model for a particular application [39]. Moreover, in this context, it seems reasonable to state that we need to opt for methods that combine evaluating the learning algorithm and the model selection [39]. Removing bias and sensitivity of external factors requires intensive computational procedures, for instance, nested cross-validation or "double cross" [43, 39].

Some concern arise when the proportions of classes in a data set are significantly different i.e. the data set is skewed or unbalanced. In fact, it is possible that the randomness of the construction of the splits makes them non representative of the distribution of the corresponding populations in real-life settings. Consequently, the results obtained for each fold will not reflect the true performance of the machine learning algorithm. To address this problem, we operate a stratified k-fold cross validation that would guarantee a realistic allocation of observations in each split to meet the same distribution characteristics of the initial full data set. Other well-tried variations of this method can be further investigated, such as distribution-balanced stratified cross-validation [44].

2.2.4 Machine Learning Classification Algorithms

2.2.4.1 Logistic Regression Logistic regression (LR) is a linear classifier which relies on separating a data set into subgroups belonging each to a class. Indeed, the output is categorical, more precisely binary or dichotomous [45]. The separation boundary being a straight line in the two dimensional space or a hyper plane in higher dimensional spaces conveys to this method the characteristic of being linear. As in statistics, logistic regression aims to fit and provide an interpretable mapping from the independent variables (input) to the dependent response (output) [45].

Two main differences arise between linear and logistic regression. First, the conditional mean of the outcome variable is defined as the outcome Y mean value given the independent variables' values x such that

$$E(Y|x) = \beta_0 + \beta_1 \times x \tag{2.1}$$

for β_0 and β_1 the unknown vectors to optimize, is a linear combination of independent variables that can take any real number [45]. Since we want the conditional mean curve to be plotted as an S-shaped curve with variable on y-axis between 0 and 1, a cumulative distribution can be used [45]. The logistic distribution is chosen for its mathematical flexibility and clinical interpretability and has the equation

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 \times x)}{1 + \exp(\beta_0 + \beta_1 \times x)} \tag{2.2}$$

[45]. The logit transformation comes then to get us back to the familiar form used in linear regression

$$f(x) = \ln(\frac{\pi(x)}{1 - \pi(x)}) = \beta_0 + \beta_1 \times x \tag{2.3}$$

and thus allows taking advantage of the properties already established for the latter [45]. Second, while the conditional distribution is normal for linear regression, it is binomial with probability of success equal to the conditional mean [45].

If we define the sigmoid function (Figure 6) previously established for the conditional mean as follows:

$$h_\beta(x) = \frac{1}{1 + \exp(-\beta^T x)} \tag{2.4}$$

mainly having the sigmoid function at its origin, applied to $\beta^T x$, we can define a function dependent on the unknown vector beta which will be optimized to fit the model. The optimization problem lays on maximizing an expression, called the log likelihood function:

$$L(\beta) = \sum_{i=1}^{n} (y_i ln(\pi(x_i)) + (1 - y_i) ln(1 - \pi(x_i))) \tag{2.5}$$

for x_i a single observation corresponding to the i^{th} component of the input vector x and y_i the i^{th} label of the outcomes vector Y [45]. The optimization algorithm used can be Liblinear, LBFGS or SAGA, among others.

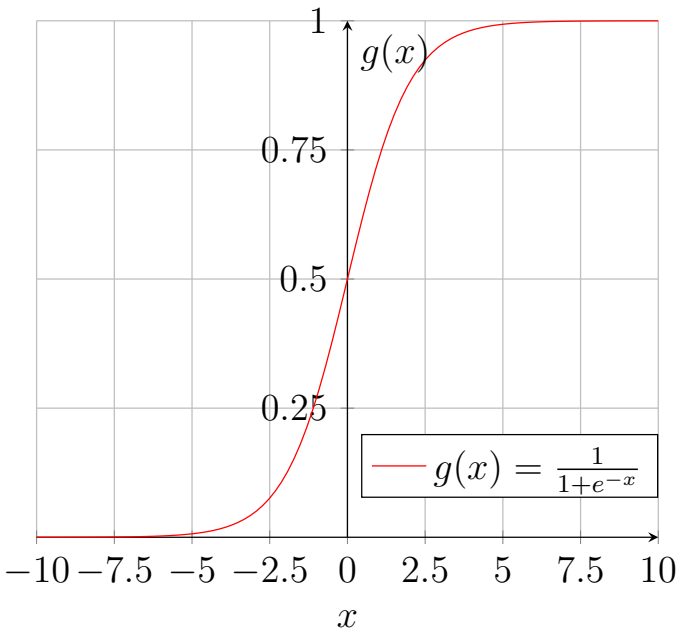

Figure 6: Plot of the sigmoid function

Finally, since the output of the sigmoid function is bounded ranging from 0 to 1, a threshold, equal to 0.5, is set to split this range into two intervals. The choice of this cutoff value assumes equal probabilities of outcome classes i.e. if the conditional mean is higher than 0.5 the output is set to be 1, it is considered zero otherwise. If the data is unbalanced (or skewed), weights can be adjusted to account for this problem.

2.2.4.2 Artificial Neural Network [46] An artificial neural network (ANN) configuration is inspired by the human network of neurons and synapses' mechanism [47]. ANN correspond to a sophisticated algorithm capable of catching underlying characteristics of a non-linear relationship between input and output variables through its main two steps. The architecture of an artificial neural network can be summarized in the following elements: multiple layers each having a finite number of neurons. Also, a bias is added in each layer and is linked only to the neurons of the next layer. Each neuron has at its input a linear combination of the previous layer's neurons' output, and at its output an activation function

$g(.)$ evaluated for this input. Let's call o_i^k the output of the i^{th} neuron in the k^{th} hidden layer, then its value is obtained according to this formula (see Figure 7 for a better understanding of the significance of the variables):

$$o_i^k = g(w_{0i}^k + \sum_{m=1}^{r_{k-1}} w_{mi}^k x_m^{k-1}) \tag{2.6}$$

where:

- x_m^{k-1} is the m^{th} output of the $(k-1)^{th}$ hidden layer (k-1 = 0 corresponds to the input layer);
- w_{mi}^k is the weight assigned to the connection going from the m^{th} output of the $(k-1)^{th}$ hidden layer to the i^{th} neuron of the k^{th} hidden layer;
- w_{0i}^k is the weight assigned to the connection going from the unit bias of the $(k-1)^{th}$ hidden layer to the i^{th} neuron of the k^{th} hidden layer;
- r_{k-1} is the number of neurons of the $(k-1)^{th}$ hidden layer.

The activation function's goal is to map the output value in the [0,1] interval corresponding to a probabilistic output range. Multiple examples of functions can be used: sigmoid function (Figure 6), hyperbolic tangent function (tanh), softmax function, softsign and rectified linear unit function, among others [48].

Two main steps are mandatory in the process of learning of an ANN:

First, the forward propagation consists of computing all the outputs of the neurons. The first layer represents always the input variables of the ANN model, plus the bias. A simple model would include an input layer (leftmost layer), a hidden layer and an output layer. However, multiple hidden layers can be included and this basic architecture can be further enhanced in the context of deep learning to form recurrent or convolutional neural networks among other structures [49]. At the end of this phase, the output layer's neurons have each calculated a value for the output.

Second, during the backpropagation phase, an error function is first determined to reflect the difference between the desired and the predicted output, the mean squared error is used. Since fitting the model is completely satisfied when finding the set of weights between the neurons of two distinct layers, an optimization algorithm is at the center of this process to

minimize the error function, which derivatives with respect to the weights are used to update the weight values. Gradient descents examples used are Adam, Stochastic Gradient Descent and LBFGS (a variation of quasi-Newton method).

The training observations are successively fed into the neural network, possibly multiple times, defining the number of epochs. The stop condition, denoting the convergence of the optimizer, can be defined by a maximum number of iterations or a minimum error improvement.

Figure 7 details the structure of an artificial neural network with multiple outputs.

An artificial neural network presents the advantage of allowing *structural stabilization* which consists of modifying the complexity of the model through varying the adaptive parameters' number [50]. Several ways con achieve this goal ranging from comparing the performances of distinct models (with unequal number of units), through starting with a wide network and eliminating the least relevant weights or units, to, on the opposite, starting with a limited network and supplementing it with units [50].

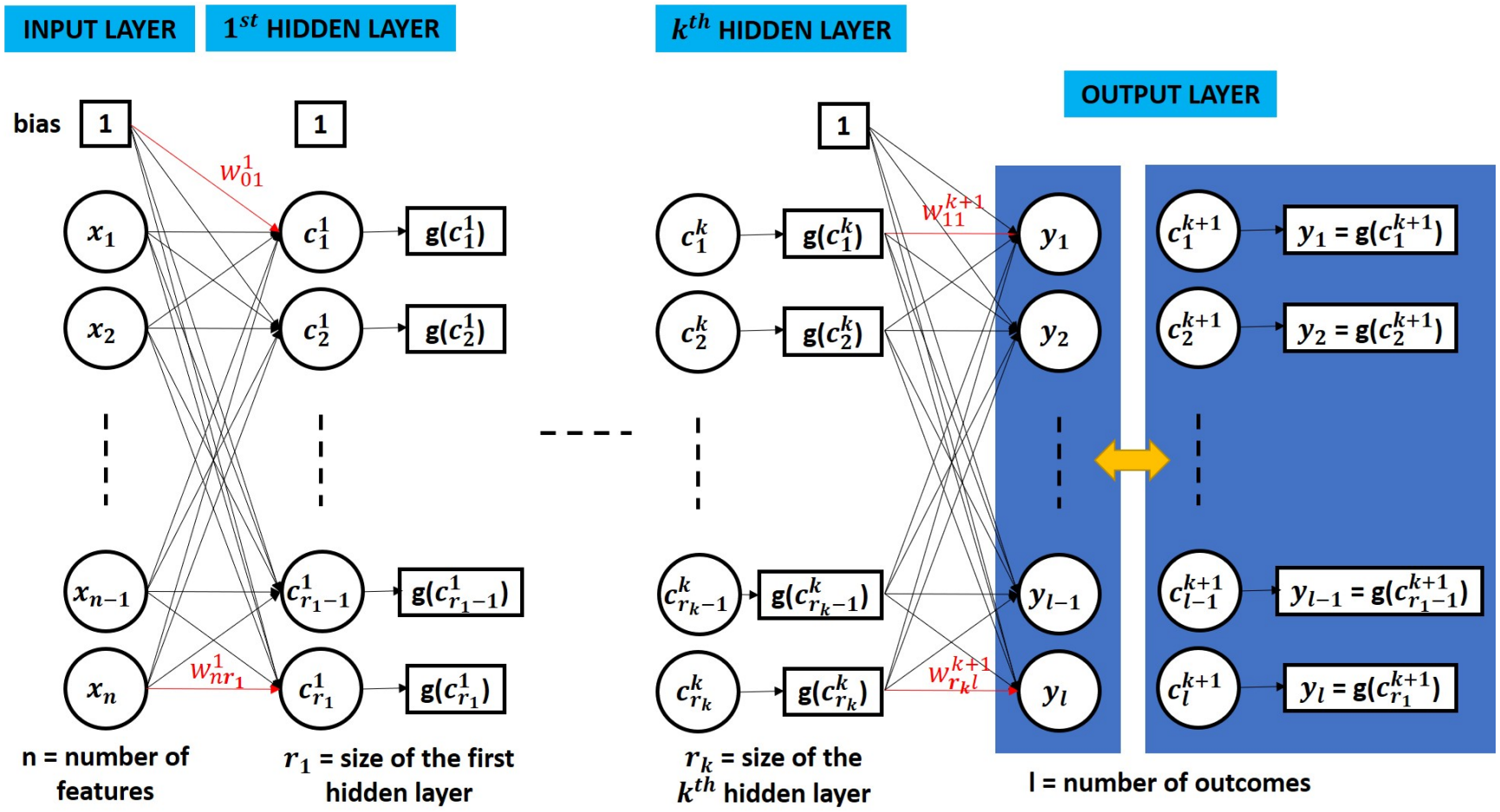

Figure 7: Structure of an artificial neural network comprising an input layer, k hidden layers and multiple outputs

2.2.5 Dimensionality Reduction

The term curse of dimensionality is employed when the dataset features are so numerous that they impact negatively the performance of the algorithm i.e. the complexity and redundancy of certain elements result in the decrease of the accuracy and the occurrence of overfitting [51]. Consequently, reducing the dimensions of a datasets to its relevant features would not only address the previously mentioned challenges but would also improve the computation time needed to run the algorithms [51]. In the absence of this pre-processing step, the number of datapoints needed to obtain a good accuracy should be significantly high to account for the sparsity of the data, the critical dimension is defined as the minimum number of dimensions that would fulfill the objective of high accuracy [51]. Dimensionality reduction algorithms are also essential to the case where one needs to visualize the data with a specific number of dimensions and lay under one of the following categories [51].

Feature Evaluators and Feature Ranking Algorithms: Algorithms that perform these tasks select significant features through a predetermined process, they are also called Filters and Wrappers [51]. In the case of Filters, the predictive efficiency of the approaches is lower than Wrappers since they perform the features assessment and scoring without taking into account the classification model that will be used [51]. Some examples to measure feature relevance include Pearson correlation coefficient, information gain and chi-squared score [51]. For Wrapper methods, error rates for subsets of features are determined using the prediction model that is going to be used for the classification, thus it leads to a more suited feature selection for the particular model [51]. A trade-off between accuracy and execution time is to be made when choosing the most convenient approach because filters, whether they are univariate (processing a variable at each iteration) or multivariate, are less computationally expensive [51].

Whereas some algorithms reduce dimensionality by forming new features as linear combinations of the original ones, other non-linear methods preserve the non linear relationships between the dimensions to be loyal to the real structure of the data, leading to a better classification performance [51]. Clustering algorithms prove efficient in the cases of supervised and unsupervised learning but fail to generalize to non-linear complex data sets [51].

Figure 8 displays some dimensionality reduction techniques.

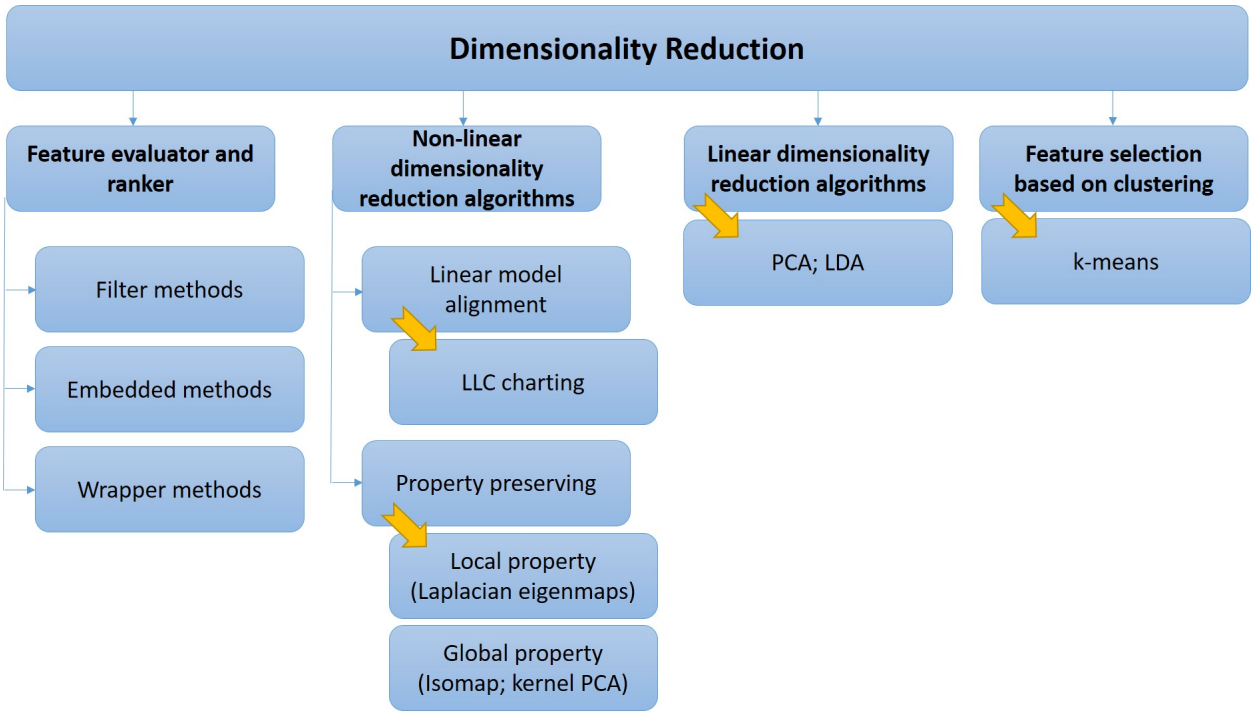

Figure 8: Dimensionality reduction techniques. Adapted from [51].

2.2.5.1 Cohen's d Effect Size Cohen's d effect size measures how distinguishable are two distributions with respect to a feature by reflecting the distance between the means. The larger the effect size, the greater the ability of the corresponding feature to separate between the curves. The dataset features should meet two assumptions relative to the classes' distributions; normality of the distributions and the homogeneity of their variances.

The effect size d is calculated using this formula:

$$d = \frac{|\overline{x_1} - \overline{x_2}|}{SD_{pooled}} \tag{2.7}$$

for $\overline{x_1}$ is the mean of the first group sample, $\overline{x_2}$ is the mean of the second group sample and SD_{pooled} is the pooled standard deviation such that:

$$SD_{pooled} = \sqrt{(\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2})} \tag{2.8}$$

for n_1 and n_2 are the sample sizes of the first and second group respectively, and SD_1 and SD_2 are the standard deviations of the first and second groups respectively.

This Effect Size can be small, medium or large for values in the interval [0, 0.2], [0.2, 0.5] and [0.5, $+\infty$].

2.2.5.2 Recursive Feature Elimination Recursive feature elimination (RFE) is a wrapper method also referred to as backwards selection [52]. The procedure is based on a succession of predictive models fed with, consecutively, the full dataset features then smaller versions of the original dataset after eliminating, at each step, a set of unimportant features [52]. The stop criterion is fixed with a minimum number of features in a dataset. This whole action is built on the ranking of the features made initially so that, at every step, variables presenting the same most inferior ranking would be removed [52]. Every classifier using a distinct reduced dataset has its performance evaluated with a measure of performance, such as the accuracy or the f-1 score. RFE selects the subset that presents the highest value of this metric and considers it as the optimal set of features. If the selected set is the original dataset, then this approach fails to operate dimensionality reduction.

Two main issues are faced when using RFE: (1) The features ranking is not updated and is statically fixed at the initialization step [52]; (2) The sequential nature of the scheme applied for features elimination inherently does not tolerate to blend features of different rankings in the same dataset [52].

2.2.5.3 Least Absolute Shrinkage and Selection Operator The least absolute shrinkage and selection operator (LASSO) paradigm relies on the reduction of the sum of squared errors between the outcome and an affine function of the input variables, while performing an L1 penalty. An optimization method is used to minimize the following function:

$$SSE_L1 = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda_L \sum_{j=1}^{P}|\beta_j| \tag{2.9}$$

[52]. This approach effectively nullifies some coefficients thus omitting the corresponding variables from the linear estimation [52]. This procedure ensures the selection of the most significant features [52]. This is the advantage of LASSO over Ridge Regression which

minimizes a similar equation but replacing the absolute value of the coefficients by their square [52]. Since it is hard to obtain a zero value of the coefficients using the latter, it is used mainly to address collinearity issues [52].

Although LASSO approach is straightforward, it requires the dataset to be standardized so that its result is not influenced by the variables scales. The regularization parameter alpha can be set by the means of a k-fold cross-validation over the provided data.

Friedman et al. (2010) [53] provided another variation of the LASSO model to tackle classification problems [52].

2.2.6 Data Collection

Data collection is at the center of any machine learning model as the input data play a deterministic role when it comes to defining the quality of the algorithm results. Indeed, the features extracted should be expressive in the sense that they should contain enough information to perform the task assigned [18]. However, it is not necessarily evident that, at the moment of the data collection, the observers harvesting the features are aware of the informative variables they need to evaluate. One way of knowing this information is to assess whether or not an expert in the corresponding field can accurately perform the machine learning task only by using the data set features [18]. Although this method is efficient, human knowledge of numerous biomedical phenomena, for instance, is lacking and we can then no longer refer to it for subsequent data gathering [18]. Consequently, it seems reasonable to save data corresponding to as much features as possible to get a complete set of observations and grasp every single detail related to the phenomena that we need to detect or describe. However, real-life phenomena are very complex and the variables that we usually choose to identify them are not necessarily independent. In fact, various features in an ECG, for example, are correlated which constitutes a redundancy in information that can be misleading to the model. In particular, a classical issue that the model can encounter is over-fitting the training data because of the small number of data points compared to the number of features [54].

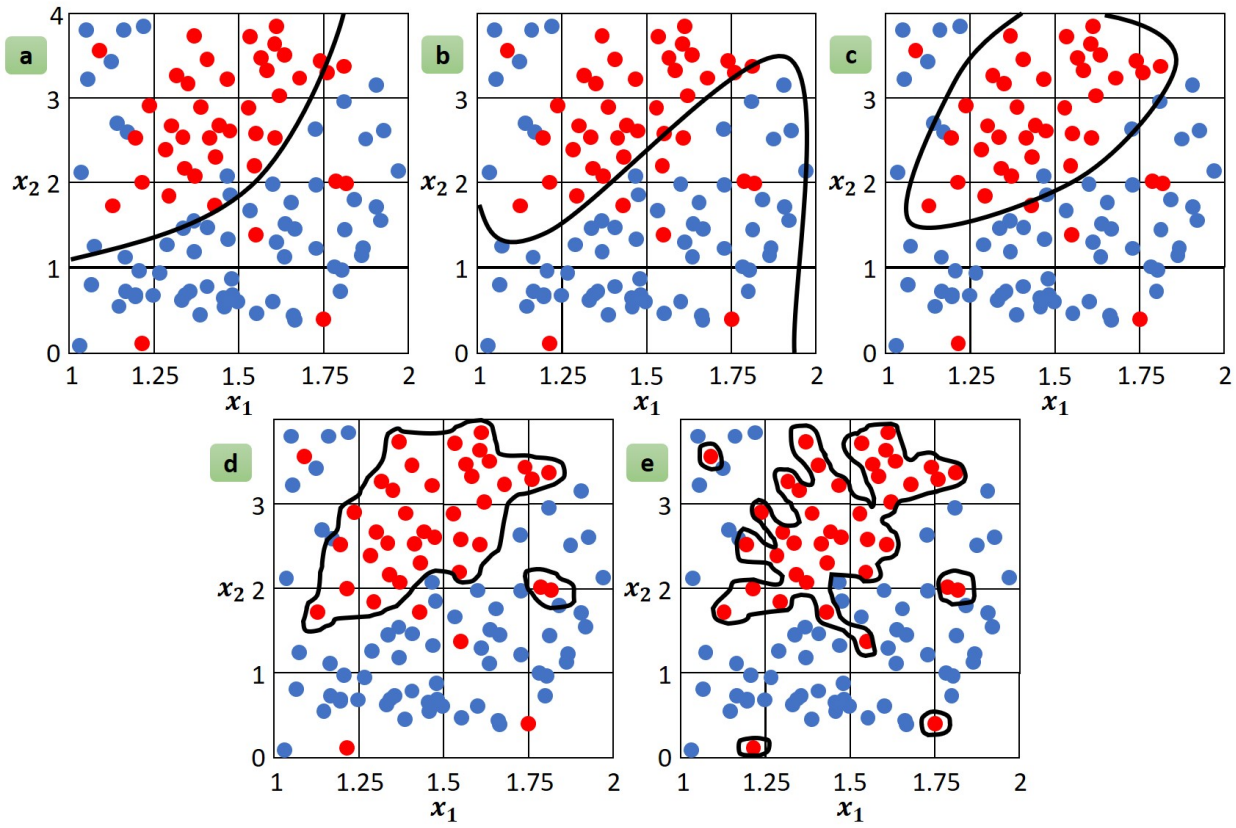

Figure 9: Illustrations of classifiers' performances on a same data set to simulate the phenomenon of overfitting, where the models in (a) and (b) are underfitting the data, the model in (c) is fitting them well and the models in (d) and (e) are overfitting them. Adapted from [39].

Overfitting (Figure 9) occurs when the *principle of parsimony*, which requires the model's content to be restricted to the minimum necessary elements to accomplish the task, is violated [55]. This phenomenon has two origins: either the model selected has more flexibility than the application calls for (e.g. using an artificial neural network to tackle a linear problem) or it involves irrelevant predictors alongside with the mandatory ones [55].

In practice, overfitting can be detected through a "poor generalization performance" characterized by a model that fails to give accurate predictions on an independent testing set while it provides good results on the training observations [54].

When a data set is high-dimensional, its usefulness in machine learning applications would be limited due to the curse of dimensionality. This term is employed to qualify the phenomenon by which a high number of variables describing each observation of a training set would require the size of the latter to increase exponentially with the features number [56]. Bishop (2006) explains this deduction by basing his reasoning on a naive approach consisting of splitting the multidimensional data set space into identical cells in each at least one training datum should be placed to confer to that cell its output label (which will be the predicted label of any testing point that will occupy that cell) (Figure 10) [56]. The *curse of dimensionality* causes the peaking phenomenon which is the observed decline in classification performance due to a low training samples to features number ratio [57].

Thus, a burden, due to the ramification of physical phenomena, is the requirement to accumulate as much data points as possible that account for the general distribution of the outcome measure and recognize all the fundamental characteristics essential to the machine learning task [54]. In spite of being an inescapable way to gain a broad knowledge during the training of the model, this tedious assignment, which may require the collection of thousands observations consisting each of the same features while respecting the above conditions on the latter, resulted in a limited predictive contribution of algorithms in clinical settings [18]. Added to the fact that this process is time-consuming and may extend to several years, it is also costly and access to certain information may be critical for certain applications due to confidentiality and privacy policies.

In statistics, power analysis is the first step to make before engaging in the experiment. This starting point fixes the minimum number of participants to recruit in order to obtain

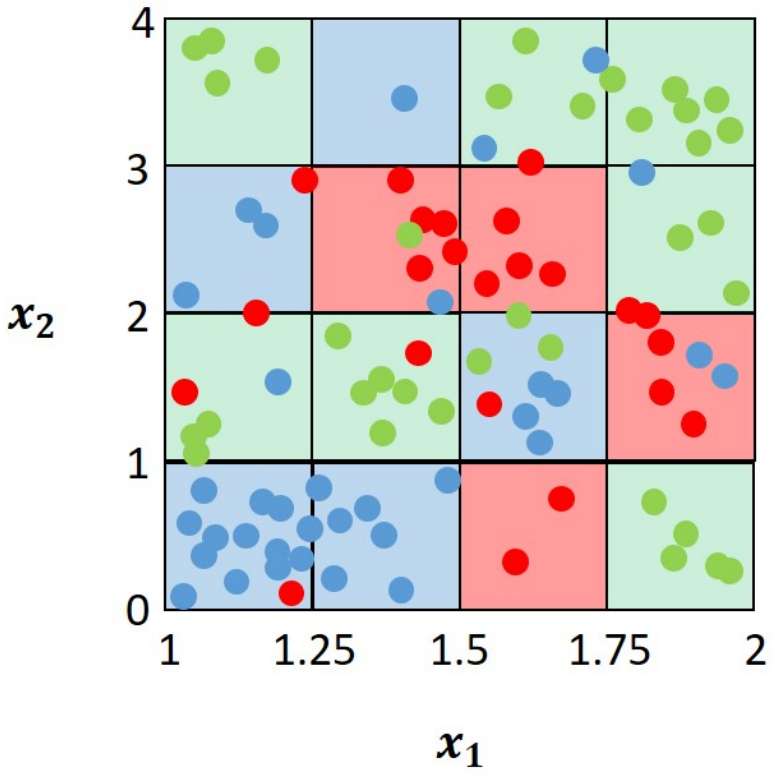

Figure 10: An illustration of an approach useful to classify data belonging to three classes by dividing a two-dimensional space into equal squares each having the label corresponding to the outcome of the majority of training points in it. Adapted from [56].

results at specific levels of Power and Type I and Type II errors. An analogous approach to sample size estimation was presented by Hajian-Tilaki [58] for biomedical research involving diagnostic tests. The formula set to calculate the minimum total sample size asks for a value of sensitivity or specificity obtainable through a thorough state-of-the-art, a confidence level (typically 95%), precision or maximum marginal error of the estimates and, finally, the prevalence of the cases (as opposed to controls) in the population in question [58]. This last element has been underlined in a previous work by Buderer [59]. By proceeding according to this method, a reliable minimum sample size value is computed allowing for a good anticipation of the experiment resources.

2.2.7 Dealing with Data Missingness

As per the definition provided by Little and Rubin (2019) in their book: "Missing data are unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value." [60]. In statistics, some cases of absence of data correspond to this definition while others don't [60]. Little and Rubin (2019) evoke the example of non response in an opinion poll explaining how missing data corresponding to the people willing to vote but do not want to reveal their choice, for instance, can be imputed while the ones of people who are not going to vote do not fall under the above definition [60].

On-field experiments may encounter several incidents resulting in loss of data, noise in recordings (contamination) or lack of certain measurements due to apparatus failure [61]. Such partial data deficiency is called item nonresponse, as opposed to the concept of wave nonresponse in the case of longitudinal research where an individual participates in multiple phases of the experiment and may miss a wave or in the extreme case stop the experiment (attrition) [61].

It is almost impossible to conduct the perfect data collection. Thus, it was inevitable to search for techniques to address this issue without altering the quality of the data or misleading the analysis results. Several methods are available to this end and include mean substitution, adding a dummy variable for missingness and regression-based single imputation [62]. A challenge to data imputation is that the estimates of the population mean or variance, and more broadly the population characteristics, can be biased i.e. significantly different from the actual values [62]. More sophisticated methods were suggested by Graham (2009) in [62] and may, for some, tackle this problem.

Finally, we can adapt the data collection procedures to take advantage of the effectiveness of handling missing data by implementing creative study designs such as the $3-form\ design$ (Figure 11) which grants 33% more questions to ask to the participants without increasing the number of questions to ask for each of them [62].

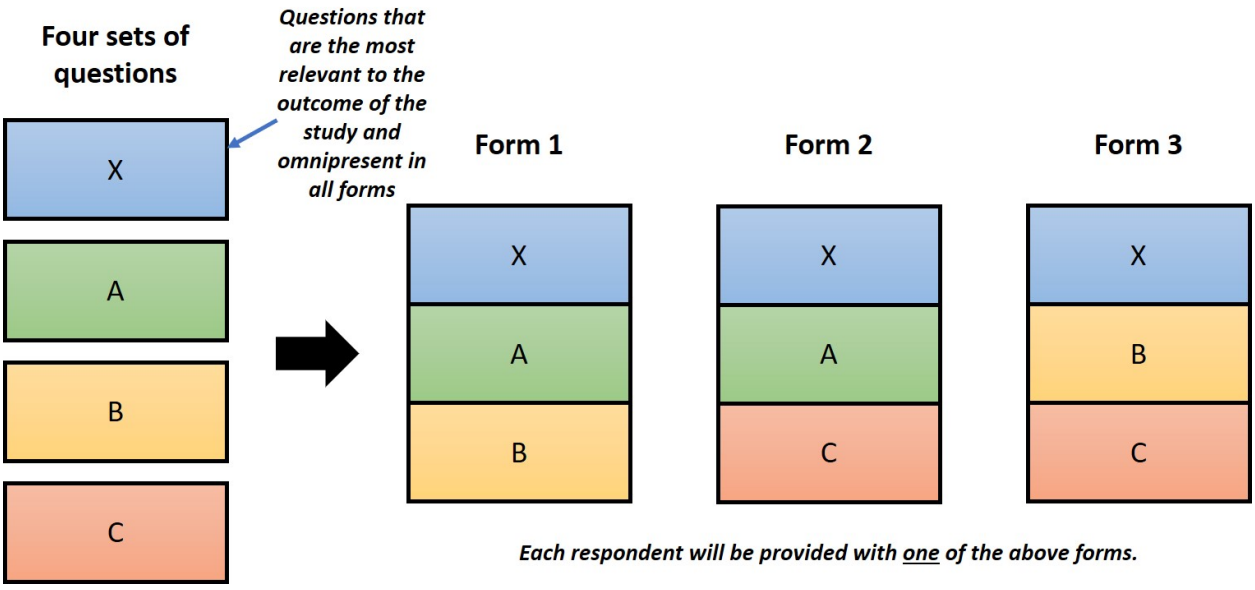

Figure 11: 3-form design principle. Illustrated as described in [62].

2.2.8 Performance Metrics

Numerous performance metrics can serve to clarify the tendency and behavior of the machine learning algorithms, among which the accuracy is the most commonly used metric. However, in certain cases, giving the percentage of correct predictions turns out to be unable to reflect the efficiency of an algorithm and how well it fulfills the role its designer created it for. For instance, if the data is imbalanced, with a large prevalence of healthy patients, a machine learning model classifying all patients as healthy would present a relatively high accuracy while being completely useless.

To avoid such misleading methods, a classical performance metric for machine learning biomedical applications is the area under the receiver operating characteristics (ROC) curve [63]. This tool is powerful because it reflects the ability of binary classifiers to distinguish between two populations (e.g. sick individuals and healthy individuals) while providing results that are not influenced by the distribution of the two classes [63]. The best performing model has the area under curve (AUC) with a value closest to 1.0.

Additionally, the confusion matrix can be used to draw pertinent conclusions. This matrix displays four numbers: the quantity of true positives, false positives, false negatives and true negatives obtained by our classification model by comparing the actual labels of our data points and the predicted labels (Figure 12). By convention, specifying the studied condition separates the sample into two populations: positive individuals who present the condition (having the number '1' as label) and negative individuals who do not present the condition (having the number '0' as label). However, some algorithms give a continuous probability at their output and a cutoff value must be fixed to return to this binary scheme. By definition, a true negative is a negative case classified as negative, while a false negative is a positive case predicted negative. The same reasoning applies to true positives and false positives.

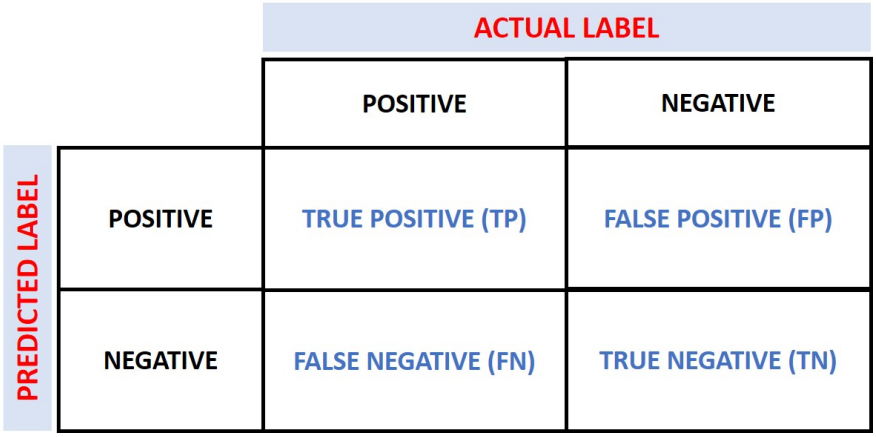

Figure 12: Confusion matrix

Using the values of the confusion matrix cells, we can define the following metrics:

- **Accuracy (ACC):** reflects the proportion of correctly predicted cases among the total number of available cases according to the equation:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.10}$$

- **Positive Predictive Value (PPV):** reflects the proportion of true positives among the predicted positive cases according to the equation:

$$PPV = \frac{TP}{TP + FP} \tag{2.11}$$

- **Negative Predictive Value (NPV):** reflects the proportion of true negatives among the predicted negative cases according to the equation:

$$NPV = \frac{TN}{TN + FN} \tag{2.12}$$

- **Specificity (SP):** reflects the proportion of true negatives among the actual negative cases according to the equation:

$$SP = \frac{TN}{TN + FP} \tag{2.13}$$

- **Sensitivity (SE):** reflects the proportion of true positives among the actual positive cases according to the equation:

$$SE = \frac{TP}{TP + FN} \tag{2.14}$$

- **F1 score:** reflects a compromise between Positive Predictive Value (precision) and sensitivity (recall):

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{2.15}$$

The above mentioned ROC curve plots the sensitivity (True Positive Rate) versus the False Positive Rate (1-specificity) for all cutoff values (Figure 13).

The ROC curve is useful in clinical medicine, beyond the only reading of its AUC [64]. For instance, it can be used to pick an optimal threshold for a sensitivity-specificity trade-off with one of the following procedures: Youden Index J and closest top-left criterion [65]. Indeed, these methods can be summarized as follows:

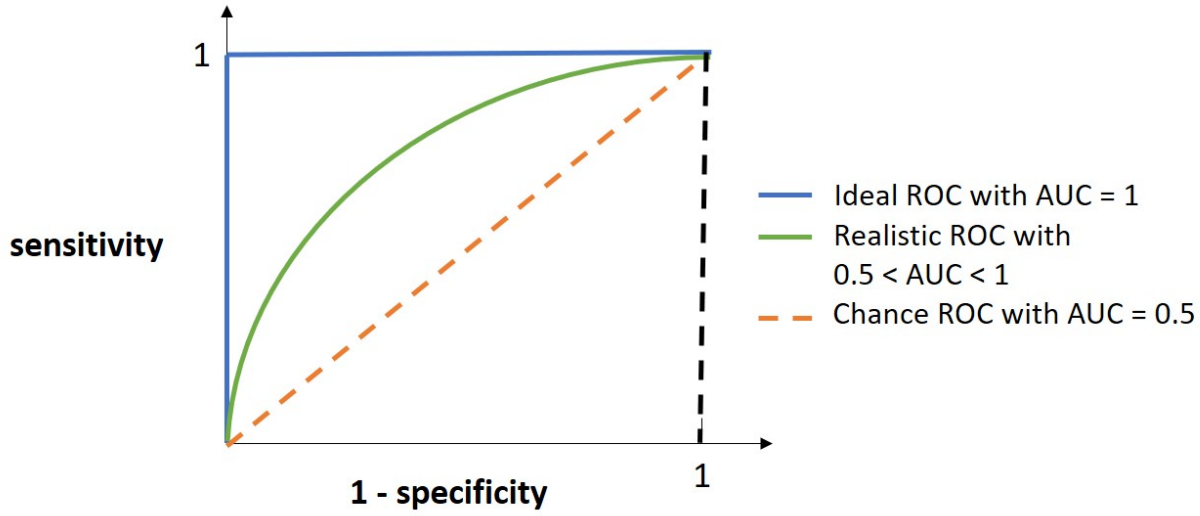

Figure 13: ROC curves with their corresponding area under curve

- **Youden Index J:** Youden Index [66] is bounded by 0 and 1, representing perfect discrimination between the two populations studied and "complete overlap", respectively [67]. This deduction finds its root from the defining equation present in [67]:

$$J = \max_{c} \quad [SE(c) + SP(c) - 1] \tag{2.16}$$

  where SE corresponds to the sensitivity, SP corresponds to specificity and c belongs to the set of values of all possible thresholds [67].

- **Closest top-left criterion:** By this method, we choose the cut-off value corresponding to the closest point to the top-left corner of the ROC curve's two-dimensional space [68] via solving a minimization problem stated in [68]:

$$q_s = \min_{c} \quad [(1 - SE(c))^2 + (1 - SP(c))^2] \tag{2.17}$$

  where the equation has been adapted from the above cited reference to conform with the notation adopted for the Youden Index.

  It is also relevant to consider other criteria like "a fixed level of specificity" [65].

3.0 Methods

3.1 Design and Settings[1]

This work is accomplished in the framework of the EMPIRE study, which is a prospective observational cohort study recruiting consecutive non-traumatic chest pain patients carried by emergency medical services to one of three UPMC-affiliated tertiary care hospitals (UPMC Presbyterian, Mercy, and Shadyside). Standard 10-second 12-lead ECGs are collected, at first medical contact, for patients suspected to have ACS in accordance with prehospital medical protocols. In the case of a high suspicion of cardiac ischemia at this initial paramedical assessment, the patient's ECG was passed to UPMC medical command. The raw digital ECG data are then constantly stored there. The participants in this study, who are all patients with the above mentioned characteristics, were followed up until 30 days after their discharge to account for their exact diagnosis while gathering the outcome data.

The full data set used for this analysis comprises 1251 available patients of the EMPIRE study. We conducted a minimum sample size estimation using the methods described by Hajian-Tilaki [58], this theoretic computational step is recommended before data collection, or, at last, before starting an experiment to check whether the recruited number of patients for the study allows for an appropriate diagnostic tests' AUC analysis. Parameters are fixed as follows: a maximum marginal error of estimates (precision) of 5%, a 95% confidence level and desired sensitivity and specificity validation values of 90%. These specifications yielded a minimum data set size to achieve ACS detection of 927 patients, considering a minimum prevalence of 15%.

The primary outcome of the study was the presence of ACS (myocardial infarction or unstable angina) during the primary indexed admission. According to the 4th Universal Definition of Myocardial Infarction guidelines, the definition of ACS consists of the pres-

[1] Portions of this section are taken from a forthcoming paper cited, in its current status, as: **S. S. Al-Zaiti, L. Besomi, Z. Bouzid, Z. Faramand, S. O. Frisch, C. Martin-Gill,R. Gregg, S. Saba, C. Callaway, and E. Sejdic, "Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram,"** ***Nature Communications*****, 2020, in press.**. (Available in reference [6] (submitted, under review).).

ence of symptoms of ischemia (i.e. diffuse discomfort in the chest, upper extremity, jaw, or epigastric area for more than 20 minutes) and at least one of the following criteria: (1) elevation of cardiac troponin I ($> 99^{th}$ percentile) with or without subsequent development of diagnostic ischemic ECG changes during hospitalization, (2) imaging evidence of new loss of viable myocardium or new regional wall motion abnormalities, or (3) coronary angiography or nuclear imaging demonstrating $> 70\%$ stenosis of a major coronary artery with or without treatment [20]. Data annotation was accomplished by two independent reviewers, with the intervention of a board-certified cardiologist in case of animosity, on the basis of not only serial ECGs but also results obtained from further extensive cardiac diagnostic tests such us echocardiography or biomarkers after running lab tests, alongside with relevant information relative to past medical records and medications intake. The principle guiding this labeling process is that a patient is declared as healthy (ACS negative) as long as adverse events are absent up until 30 days of follow-up.

A summary of the different steps of the study design is provided on Figure 14.

3.2 Data Preprocessing[2]

The manual over-reading of the ECGs revealed the ones with excessive noise or artifact, which were excluded (including ECGs of patients with ventricular tachycardia or fibrillation) or replaced by the next serial ECG obtained from the emergency assessment. The patients removed from the study represent almost 0.5% of the total number of patients in the data set (Cohort 1: n=5/750; Cohort 2: n=2/501). We included the rest of the accessible ECGs in our work, counting those with pacing, bundle branch blocks, atrial fibrillation, or left ventricular hypertrophy.

[2]Portions of this section are taken from two forthcoming papers cited, in their current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** and **S. S. Al-Zaiti, L. Besomi, Z. Bouzid, Z. Faramand, S. O. Frisch, C. Martin-Gill,R. Gregg, S. Saba, C. Callaway, and E. Sejdic, "Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram,"*Nature Communications*, 2020, in press.**. (Available in references [1] (submitted, under review) and [6] (submitted, under review).).

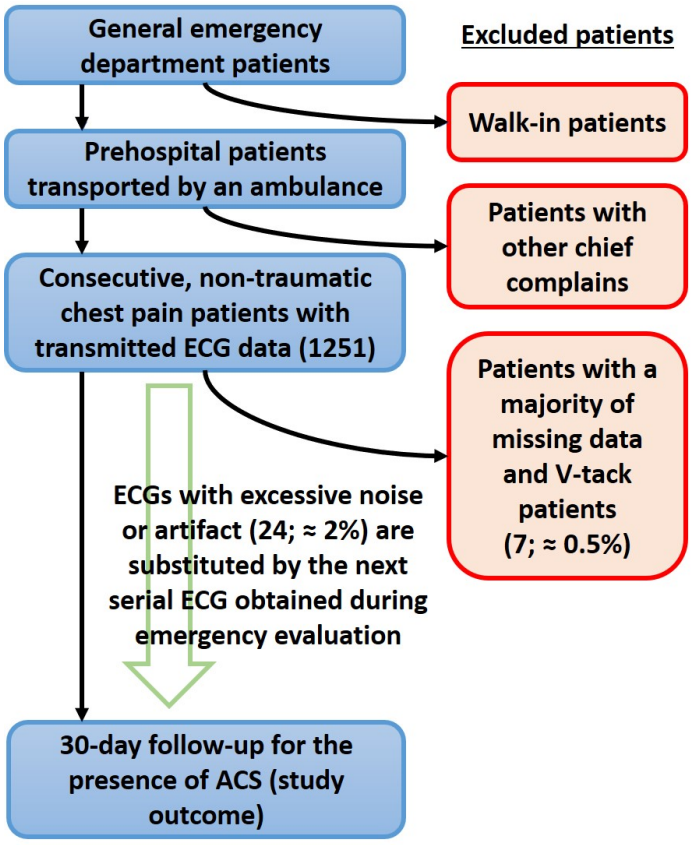

Figure 14: Study design

Afterwards, a manufacture-specific commercial software, designed at Philips Healthcare Advanced Algorithm Research Center (Andover, MA), preprocessed the 10-second, 12-lead ECG signals (500 s/s, HeartStart MRx, Philips Healthcare). Decompressing raw ECG signals resulted in the extraction of ECG leads. Next came the removal of noise, artifact, and ectopic beats, and the computation of representative average beats for every ECG lead in order to cancel remaining baseline noise and artifacts. We achieve, following these steps, a high signal-to-noise ratio and a stable average waveform signal through all 12 leads.

After their de-identification (labeling with study ID) and secure storage, all 12-lead ECGs were classified by physicians blinded to the actual diagnostic output, and using the aforementioned supplemental evaluation tools. The labeling of the diagnostic ECG changes was operated according to the fourth Universal Definition of Myocardial Infarction consensus

statement [20] as two contiguous leads with (1) ST elevation in V2–V3 $\geqslant$ 2 mm in men $\geqslant$ 40 years, $\geqslant$ 2.5 mm in men $<$ 40 years, or $\geqslant$ 1.5 mm in women; or ST elevation $\geqslant$ 1 mm in other leads; (2) new horizontal or downsloping ST depression $\geqslant$ 0.5 mm; or (3) T wave inversion $>$ 1 mm in leads with prominent R wave or R/S ratio $>$ 1, and considering other ECG findings indicative of possible ischemia, which are contiguous territorial involvement, evidence of reciprocal changes, changes beyond those caused by secondary repolarization, and lack of ECG evidence of non-ischemic chest pain etiologies.

A parallel process was started and consisted of features extraction. In fact, the amplitude, duration, and/or area measures of the P wave, Q wave, R wave, S wave, qR wave, rS wave, QRS complex, QRS peak, ST segment, T wave, STT wave, QT interval, PP interval, RP interval, and SP interval were calculated from the representative beats (k=384). Furthermore, the amplitude of ST onset, ST peak, ST offset, and J+80, along with ST slope were derived from the leads (k=60). Thus, a total of 444 temporal ECG features was obtained (Figure 15A). A group of global measures was then computed including QRS, JTend, JTpeak, Tpeak-end, and QT interval measures (k=6); QRS and T axes from the frontal, horizontal, and XYZ planes (k=16); spatial angle between QRS and T waveforms (k=6); inflection, amplitude, and slope of global QT, QRS, and T wave in frontal and horizontal planes (k=56); ratios between PCA eigenvalues of QRS, STT, J, and T subintervals (k=13); T wave morphology and loop (k=7); signal noise values (k=6); regional myocardial infarction scar using Selvester score (k=19); and injury vector gradient and amplitude (k=14). All of these numbers summed up introduce 143 spatial ECG features to our data set.

Finally, 587 temporal-spatial features are acquired from a single ECG. A manual evaluation was accomplished on every patient's record to check the ECGs' quality due to potential, relatively frequent malfunctions in prehospital settings such as electrodes' misplacement or improper sticking on the patient's chest. Presumably, the fact that the collected ECGs are only the ones communicated to medical centers, a minor percentage of ECGs suffered from an excess of noise ($<$ 3%). An in-depth exploration of the data set revealed the presence of features (n = 33) with a high prevalence of zeros across patients ($<$ 5% of non-null values). Clinicians concluded that, taking into account their electrophysiological significance, these

features should be normally null because of, for example, the absence of S waves in leads II, aV_L, V_5 and V_6, and Q waves in the majority of the leads. Thus, it was relevant to get rid of these variables, which yielded a final data set consisting of 554 features.

Figure 15 summarizes major techniques leading to this final data set. Figure 15A corresponds to calculations of durations, amplitudes and areas of various waveform deflections (444 temporal ECG features). Figure 15B shows the superposition of 12 beats each representing a lead with the calculation of global intervals and sub-intervals (6 supplemental temporal ECG features). Figure 15C presents waves resulting from the application of principal component analysis on time-voltage data of orthogonal leads I, II, V_1–V_6 in order to generate ratios of the eigenvalues corresponding to different ECG waveforms (13 spatial ECG features). In the end, Figure 15D clarifies the concept behind the estimation of axes, angles, loops and gradients of QRS and T vectors from xy, xz, yz, and xyz planes (91 supplemental spatial ECG features).

We performed a z-score normalization on the mined features and imputed missing data ($< 0.2\%$ of the total data set values) using the mean value of the corresponding features as the latter exclusively consisted of continuous variables.

3.3 Feature Selection Using Data-Driven Models[3]

We used three different data-driven algorithms to identify a list of features that were most important for optimizing the performance of the classification algorithm. First, we used Cohen's d effect size, which compares how distinguishable ACS versus non-ACS distributions of a given feature are in terms of the distance between the means. All distributions were evaluated for normality of distributions and homogeneity of variances. Features corresponding to an effect size lower than 0.35 are assumed to fail to differentiate between the two populations and were excluded from our dataset. Using this cutoff value, only 34 features

[3]This section is taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"** ***Journal of American Heart Association*****, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

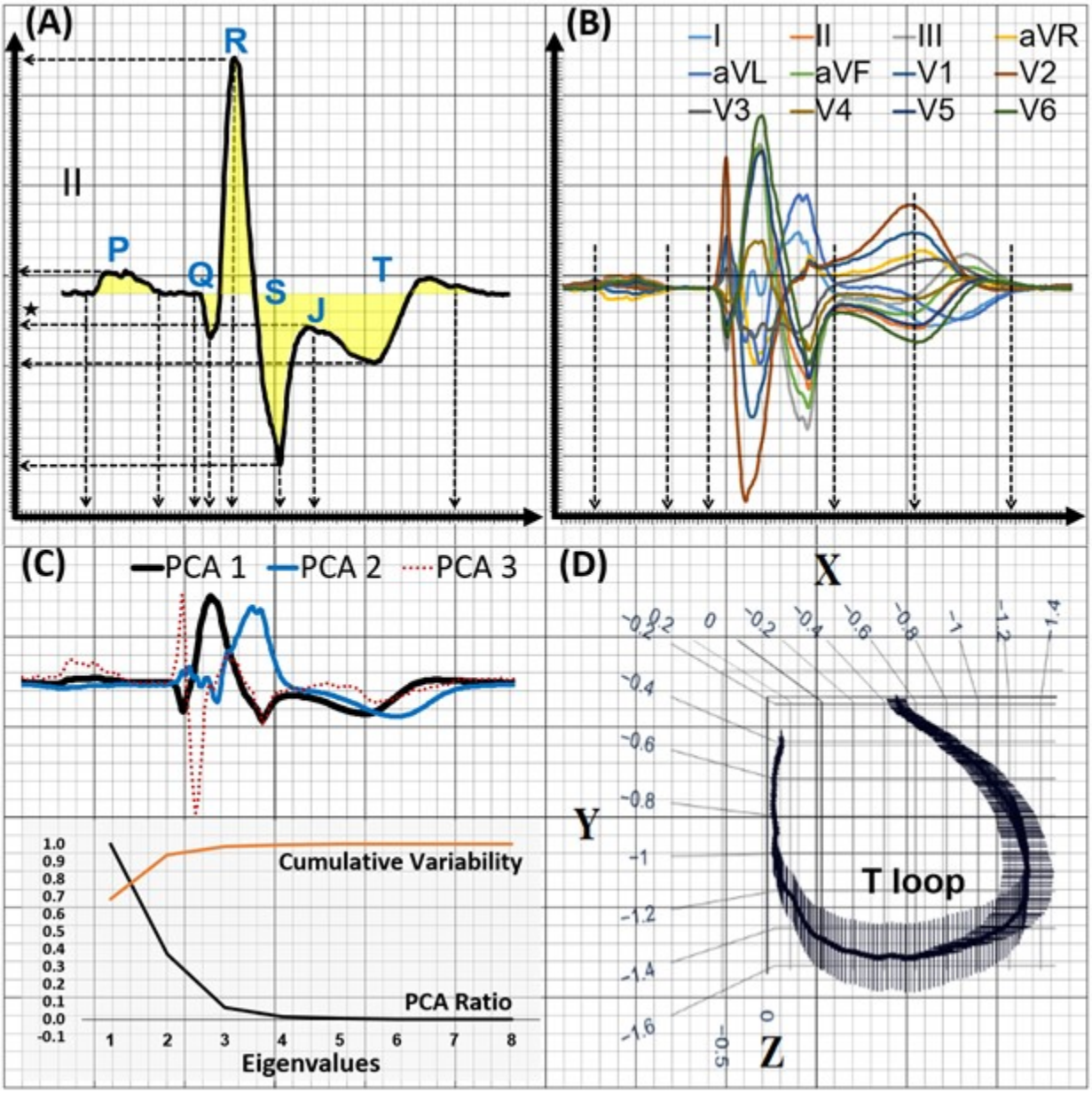

Figure 15: Illustration of the computation of 554 features from each 12-lead ECG

out of 554 remained (6%). Second, we used recursive features elimination as part of logistic regression. We evaluated 20 features per iteration and used F1 scores to evaluate model performance. F1 scores provides the best tradeoff between precision and recall using imbalanced datasets like ours, which had a 6:1 ratio of non-ACS to ACS subgroups. The selection of the optimal set of features went through a 10-fold cross-validation process. Using this technique, 156 features out of 554 (28%) were selected. Finally, we used LASSO regression to select the most important features with non-zero coefficients. We standardized all features using z-scores then used the L1 norm method to penalize the least square error between the outcome and an affine function of the input variables. The regularization parameter alpha was set by the means of a 10-fold cross-validation. Using this technique, 96 features out of 554 (17%) were selected.

Next, given that the three feature selection techniques described above use complementary, non-competing approaches, we identified the features that received at least one vote (i.e., appeared in at least one feature selection algorithm). This yielded a total of 229 features. We used these data-driven features in subsequent training and testing of machine learning classifiers in order to compare against the domain-specific manually selected features.

Figure 16 summarizes the characteristics of the three data set versions resulting from the EMPIRE data set.

3.4 ACS Prediction[4]

Logistic regression and artificial neural networks have been preferentially used in previous studies focusing on ECG-based prediction of ACS [69, 70, 71]. Considering the size of our dataset and the expected reduction of model complexity achieved through feature subset selection, we started with LR as the machine-learning classifier of choice to address the aims

[4]This section is taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"** ***Journal of American Heart Association*****, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

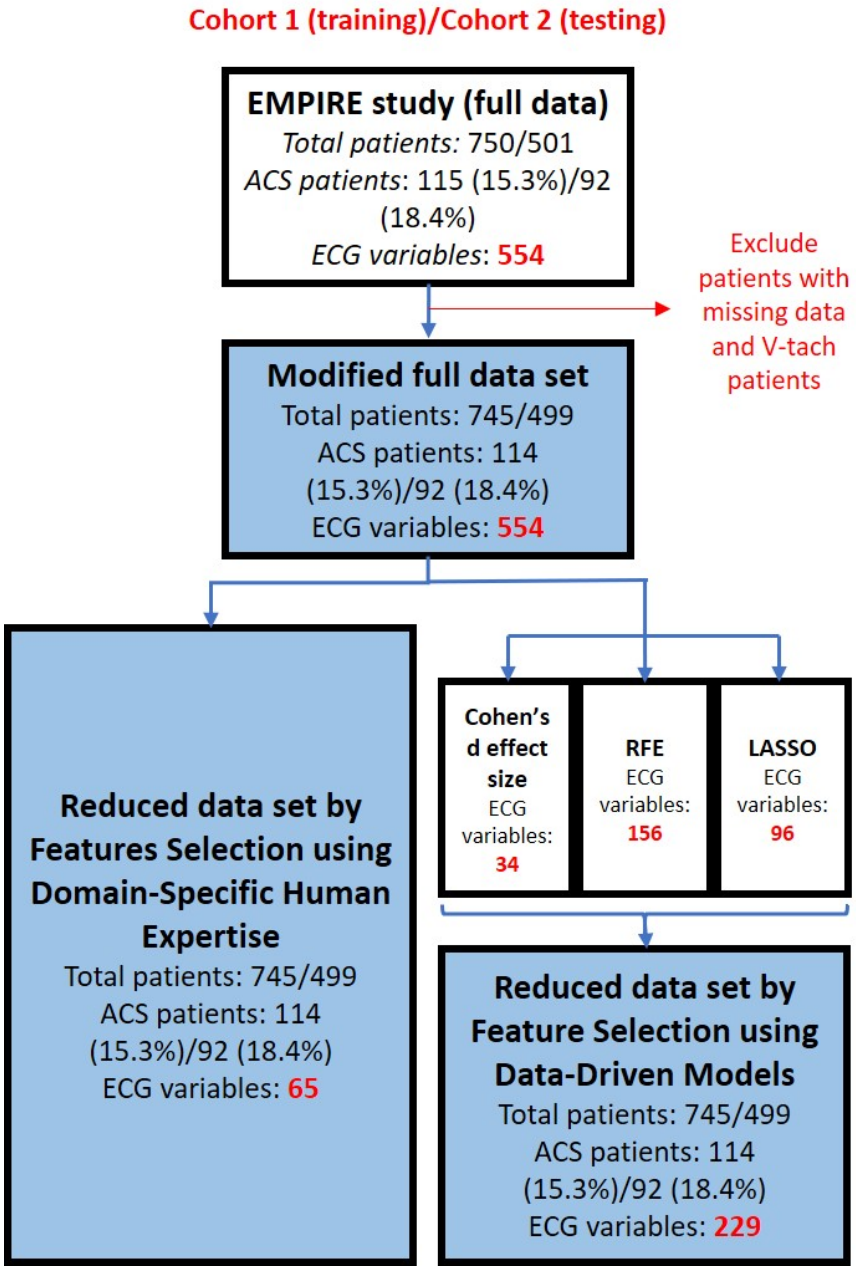

Figure 16: Schematic of the characteristics of the used data sets (in blue) originated from the EMPIRE data set

of our study. We then used ANN to explore whether feature selection approaches would have a similar effect on more sophisticated, non-linear machine learning classifiers.

Our LR and ANN classifiers were trained using a 10-fold cross-validation on Cohort 1 and, afterwards, tested on an independent Cohort 2 being completely blinded to its outcomes. We started with all 556 available features (554 ECG features with age and sex) without any feature subset selection (i.e., LR_{554} and ANN_{554}). Next, we used only the 65 manual features selected by domain-specific human experts (i.e., LR_{65} and ANN_{65}). Finally, we used the 229 data-driven features to train and test our classifiers (i.e., LR_{229} and ANN_{229}). The algorithms were trained using 10-fold cross-validation and then evaluated on an independent testing set that were blinded to the outputs.

3.4.1 Features Selection Using Domain-Specific Human Expertise[5]

Two research scientists trained in cardiac electrophysiology reviewed the 554 extracted ECG features and agreed on a reduced set of 65 features that had strong physiological basis as plausible markers of acute myocardial ischemia, including ST amplitude at J+80 and T wave amplitude from each of the 12 leads (k=24); global QRS, JTend, JTpeak, Tpeak-end, and QT interval measures (k=6); spatial axis and angles of QRS and T waves (k=8); inflection, amplitude, and slope of T wave in frontal plane (k=5); ratios between principal component analysis eigenvalues of QRS, STT, J, and T subintervals (k=13); T wave morphology and T loop features (k=7); and high frequency signal noise values (k=2).

[5]This section is taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

3.4.2 Comparison of Performance[6]

The classification performance of each classifier was evaluated using the area under the receiver operating characteristic curve. This tool is powerful because it reflects the ability of binary classifiers to distinguish between two populations. We used DeLong's test to compare the difference between the mean AUC of two correlated ROC curves of different classifiers [72], and we opted for pairwise comparisons. We set alpha at $p<0.05$ for two tailed hypothesis testing.

We also computed the accuracy, negative predictive value, positive predictive value, sensitivity and specificity of LR and ANN classifiers applied to the three data sets at training and testing. The cut-point was chosen to set the specificity at 50% and maximize the sensitivity.

[6]Portions of this section are taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

4.0 Results

4.1 Baseline Characteristics[1]

Our sample consisted of 1,244 patients from two study cohorts: training cohort (n=745, age 59 ± 17, 42% Female, 40% Black) and testing cohort (n=499, age 59 ± 16, 49% Female, 40% Black). The majority of patients were evaluated for chest pain (90%) or shortness of breathing (39%); most patients presented in normal sinus rhythm (88%) or atrial fibrillation (9%); and the rate of 30-day cardiovascular death was 4.6%. Table 1 summarizes the baseline characteristics of each cohort. The two cohorts were comparable in terms of demographics, past medical history, chief complaint, baseline ECG characteristics, and clinical outcomes.

4.2 Performance of Machine Learning Classifiers[2]

The primary study outcome, which is ACS, occurred in 114 out of 745 patients (15.3%) in the training cohort and 92 out of 499 patients (18.4%) in the testing cohort. Figure 17 shows the preliminary classification results obtained for the individual data-driven feature selection techniques, the physiology-driven technique and the full data set (no feature selection). Although the feature selection algorithms failed to generalize to an unseen testing set, they provided excellent training results. Thus, we conducted the combination procedure described in Section 3.3. Figure 18 plots the AUC of the ROC curves for all different versions of LR and ANN classifiers treated in this analysis. On training set (Figure 18A,

[1]This section is taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

[2]Portions of this section are taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

Table 1: Baseline study characteristics: Demographics and health characteristics of individuals included in this baseline study for analysis of ACS. The data comes from two cohorts - cohort 1 (n=745), used as the training set and cohort 2 (n=499), which was used as the testing set

	Cohort 1 (n=745) (Training Set)	**Cohort 2 (n=499) (Testing Set)**
Demographics		
Age in years	59 ± 17	59 ± 16
Sex (Female)	317 (42%)	243 (49%)
Race (Black)	301 (40%)	202 (40%)
Past Medical History		
Hypertension	519 (69%)	329 (66%)
Diabetes mellitus	196 (26%)	132 (26%)
Old myocardial infarction	205 (27%)	122 (24%)
Known CAD	248 (33%)	179 (36%)
Known heart failure	130 (17%)	74 (15%)
Prior PCI / CABG	207 (28%)	124 (25%)
Clinical Presentation		
Chest Pain	665 (89%)	454 (91%)
Shortness of Breathing	250 (34%)	234 (47%)
Normal Sinus Rhythm	648 (87%)	442 (88%)
Atrial Fibrillation	71 (9%)	46 (9%)
Course of Hospitalization		
Length of Stay (median [IQR])	2.3 [1.0–3.0]	1.2 [0.6-2.5]
Confirmed ACS	114 (15.3%)	92 (18.4%)
Treated by Primary PCI / CABG	74 (10%)	65 (13%)
30-Day CV Death	33 (4.4%)	24 (4.8%)

CAD: Coronary Artery Disease, PCI: Percutaneous Coronary Intervention, CABG: Coronary Artery Bypass Graft, IQR: Interquartile Range, ACS: Acute Coronary Syndrome, CV Death: Cardiovascular Death.

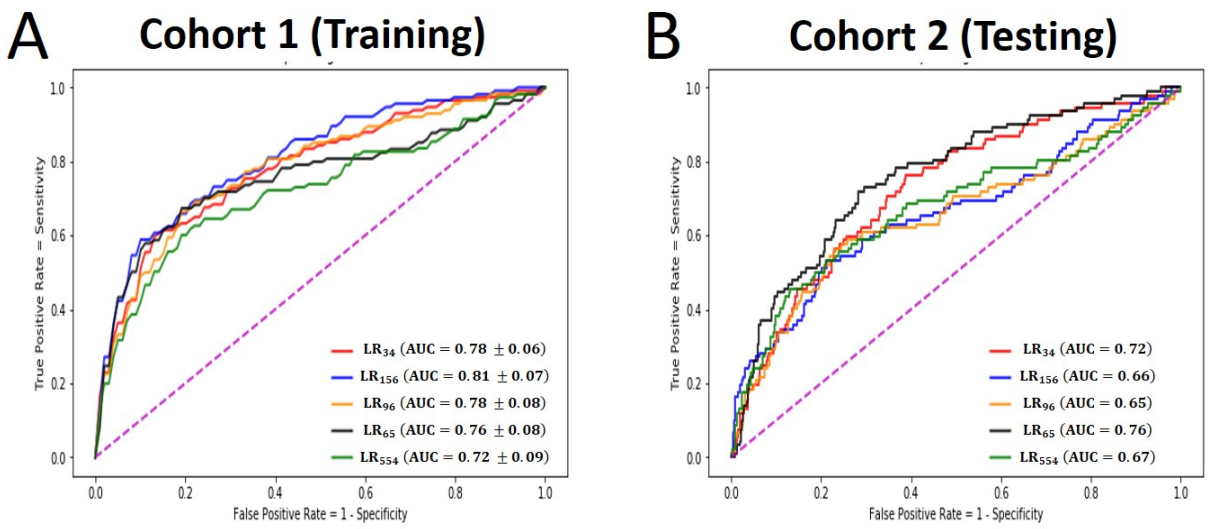

Figure 17: Preliminary results of the classification performance using LR applied to the data sets derived using Cohen's d effect size (LR_{34}), RFE (LR_{156}), LASSO (LR_{96}) and manual selection (LR_{65}) alongside the full data set (LR_{554}) on training (A) and testing (B)

left panel), both manual feature selection and data-driven feature selection techniques had better performance compared to no feature selection, with the best performance (lowest bias) achieved using the data-driven approach. However, on independent testing (Figure 18A, right panel), data-driven feature selection approach generalized poorly (high variance). Manual feature selection, on the other hand, generalized well to the testing set, suggesting a better bias-variance tradeoff. Comparing the area under ROC curve of manual feature selection and data-driven feature selection yielded a statistically significant difference for the logistic regression model with a p-value equal to 0.0105. The same trend was observed using ANN. The data-driven feature selection approach performed best on the training set (Figure 18B, left panel), but generalized poorly to the testing set (Figure 18B, right panel), again suggesting more overfitting compared to manual feature selection approach, with a p-value equal to 0.0411.

Table 2 comprises all relevant performance metrics' values, the best performance for each metric (column) is bolded for training and underlined for testing. We notice that the

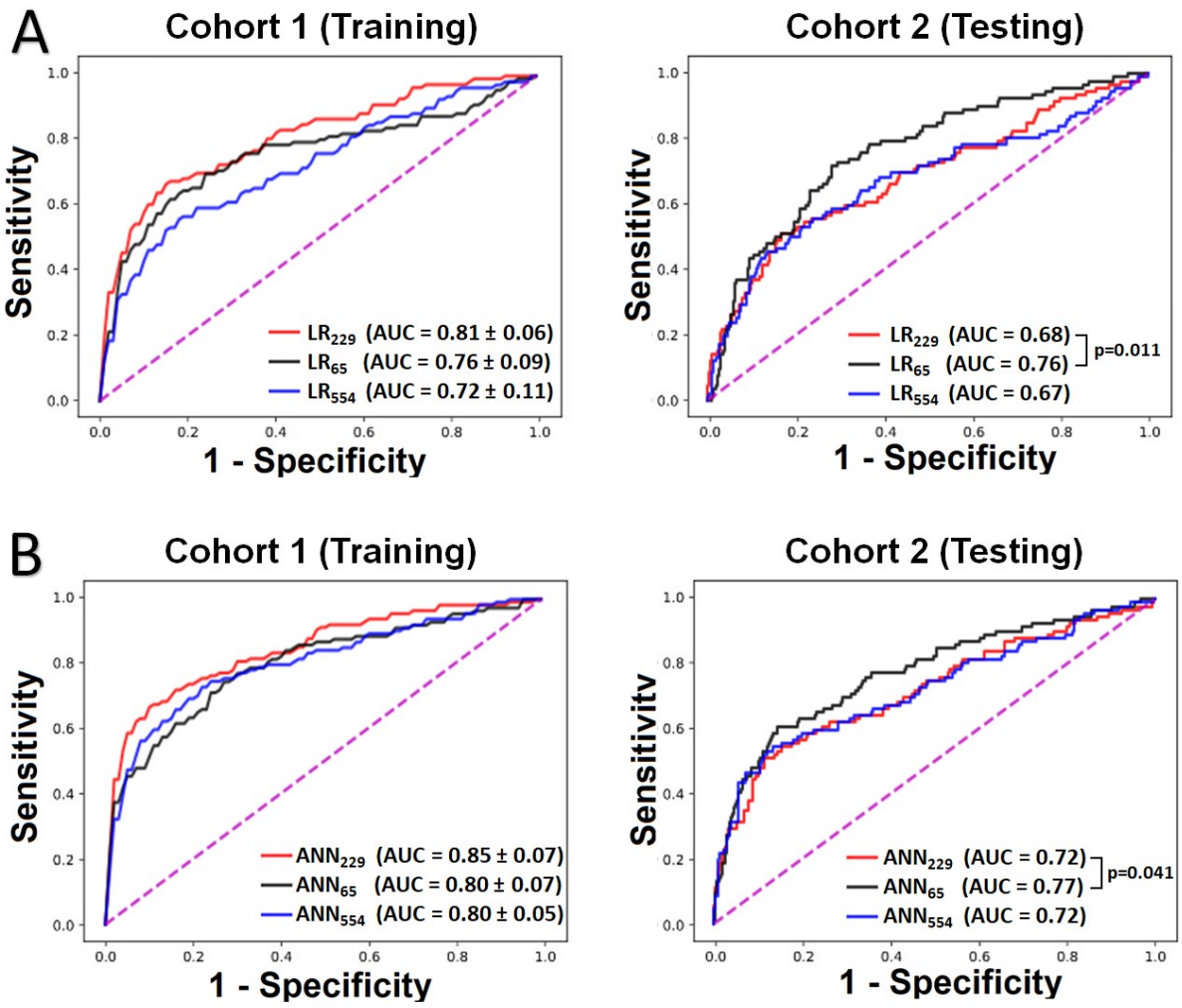

Figure 18: Classification performance using LR (A) and ANN (B) classifiers: These plots show the performance of LR and ANN classifiers on training data (Cohort 1) and testing data (Cohort 2) using all available ECG features (k=554), data-driven subset of ECG features (k=229), or physiology-driven subset of ECG features (k=65). P values are based on non-parametric method by Delong.

preeminent values for training are all obtained using the data-driven data set while the top values for testing are rather obtained with either the full data set or the physiology-driven data set. These findings also confirm that the accuracy is not the metric the maximize in our case (with 15.3% prevalence of ACS). We can reach a good predictive value (75% sensitivity and 66.34% specificity for testing on physiology-driven LR) without exceeding 68% of accuracy.

Table 2: Detailed results of the performance of the classifiers on training and testing for three versions of the data sets

Data set	Training or testing	Classifier	AUC	TN	FP	FN	TP	ACC (%)	PPV (%)	NPV (%)	SP (%)	SE (%)
Full-features	Training	LR	0.72 ± 0.11	353	278	39	75	57.45	21.25	90.05	55.94	65.79
		ANN	0.80 ± 0.05	354	277	30	84	58.79	23.27	92.19	56.10	73.68
	Testing	LR	0.67	288	119	39	53	68.34	30.81	88.07	70.76	57.61
		ANN	0.72	265	142	33	59	64.93	29.35	88.93	65.11	64.13
Physiology-driven	Training	LR	0.76 ± 0.09	338	293	30	84	56.64	22.28	91.85	53.57	73.68
		ANN	0.80 ± 0.07	345	286	28	86	57.85	23.12	92.49	54.68	75.44
	Testing	LR	0.76	270	137	23	69	67.94	33.50	92.15	66.34	75
		ANN	0.77	269	138	26	66	67.13	32.35	91.19	66.09	71.74
Data-driven	Training	LR	0.81 ± 0.06	334	297	22	92	57.18	23.65	**93.82**	52.93	**80.70**
		ANN	$\mathbf{0.85 \pm 0.07}$	363	268	24	90	**60.81**	25.14	93.80	**57.53**	78.95
	Testing	LR	0.68	254	153	36	56	62.12	26.79	87.59	62.41	60.87
		ANN	0.72	271	136	34	58	65.93	29.90	88.85	66.58	63.04

TN = True Negatives; FP = False Positives; FN = False Negatives; TP = True Positives; ACC = Accuracy; PPV = Positive Predictive Value; NPV = Negative Predictive Value; SP = Specificity; SE = Sensitivity.

4.3 Overlap in Features between Feature Selection Approaches[3]

Among the 229 data-driven features, 31 features (14%) were among the ones manually selected by human experts. These data-driven features with physiological plausibility for ACS classification included (1) lead-specific ST and T wave amplitudes; (2) T peak–Tend interval; (3) frontal and horizontal QRS and T axes; (4) spatial QRS-T angle and total-cosine R-to-T angle; (5) T loop morphology dispersion; (6) PCA ratio of QRST waveform, STT waveform, and T wave; and (7) the non-dipolar component of J wave. Among these features, T peak–T end was specifically selected by all three data-driven feature selection algorithms, and was also ranked by LR classifiers as the most important feature among the ones selected by human experts. Finally, to discern which data-driven features contributed to noise vs. contributed to true prognostic value in ACS prediction, we mapped the 229 data-driven features against the major components of the 12-lead ECG signal (Table 3). This table highlights a potential subset of features that data-driven algorithms ranked as important for the task of ACS detection but were not selected by domain-specific experts.

[3]This section is taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

Table 3: Overlap in features between data-driven and human-expert techniques

12-Lead ECG Component	**Number of Features Selected**		**Comparison between techniques**	
	Human Expert	**Data-Driven**	**Overlap in Features**	**Features Overlooked by Clinicians**
ECG normalization (k=2)	2	2	Age and sex	-
P duration, amplitude, or area (k=72)	0	25	-	Lead-specific P duration & amplitude
PR interval metrics (k=26)	1	11	Global PR interval	Lead-specific PR interval
Q duration or amplitude (k=24)	0	10	-	Lead-specific Q wave presence
R duration or amplitude (k=48)	0	23	-	Lead-specific R amplitude
S duration or amplitude (k=48)	0	16	-	S amplitude in precordial leads
Other QRS complex metrics (k=74)	1	31	Global QRS duration	QRS notch; ventricular activation time; lead-specific QRS duration or area
Selvester Score (k=19)	1	0	Total scar size	-
ST amplitude, duration, or slope (k=72)	12	31	Lead-specific ST amplitude	Lead-specific ST duration and slope
ST deviation morphology (k=14)	0	7	-	Presence of concaved ST deviation
T duration, amplitude, or area (k=76)	14	33	Lead-specific T amplitude, T-to-R relative amplitude	Lead-specific T duration or area; presence of notched T wave
QT interval and subintervals (k=23)	4	12	Global QTc, T peak-T end	Lead-specific QT interval
QRS axis (k=12)	1	7	Frontal plane QRS axis	Horizontal and spatial QRS axis
T axis (k=11)	4	6	T axis in frontal, horizontal, and spatial planes	-
QRS and T vector angles (k=5)	2	3	QRS-T angle and TCRT	-
T loop morphology (k=6)	4	4	T asymmetry & dispersion	-
Principal Components Analysis (k=16)	16	6	PCA ration of J, T, and STT	-
Noise signal (k=8)	3	2	Noise & baseline wander	-

5.0 Discussion[1]

This study assessed the efficiency of two feature selection techniques when applied to supervised machine learning classifiers, in improving ECG-based ACS diagnosis. The available data from two prospective clinical cohorts was used and led to the following ascertainment: the best bias-variance tradeoff corresponds to machine learning classifiers guided by clinical experts in the feature selection phase. We drew this conclusion from the comparison of the aforementioned classifier with no feature selection or data-driven feature selection. On a separate testing set, our study confirms the tendency observed for training in terms of the better performance of physiology-driven feature selection compared to data-driven feature selection (AUC = 0.76 vs. 0.68 for LR, and 0.77 vs. 0.72 for ANN, respectively). Besides, the manual selection based algorithms generalize better from training data (ΔAUC = 0.00 vs –0.13 for LR, and –0.03 vs. 0.13 for ANN, respectively). Also, a compelling fact is that our data analysis demonstrates the same observed effect of feature subset selection on LR (simple classifier) or ANN (sophisticated classifier).

5.1 Effect of Feature Subset Selection Approach on Classifiers Performance

Our data analysis shows that, compared to no feature subset selection, physiology-driven features optimized our LR classifier and yielded a generalizable model. This finding is expected given that using domain-specific knowledge not only tremendously reduced the dimensionality (65 out of 556 features), but also intuitively reduced the redundancy in the data, both of which are compatible with linear classifiers. On the other hand, our data analysis shows that the initial gain observed by using data-selected features generalized poorly to an independent unseen cohort. Our training set results are similar to the ones reported by

[1]Portions of this chapter are taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

Green et al. (2006) generated with 16 ECG features chosen using the principal component analysis approach [70]. Their cohort consisted of a comparable sample size (634 patients) and ACS prevalence (130 ACS patients i.e. $\approx 20.5\%$) [70]. However, Green et al. did not have an independent testing set for validation [70]. In our data, we showed that data-driven feature selection lacked generalizability on a new test example, indicating overfitting of training data coupled with a substantial variability of classifier performance. Although this finding was surprising, the small dataset size as well as the inclusion of patients with confounders in our datasets could provide a simple rationale for this unexpected finding.

We observed similar trends in results when we applied ANN as a non-linear classifier. These findings are a little bit counterintuitive given that ANN is expected to better capture the underlying characteristics of the dataset when fed with more features. This divergence can be attribute to the small sample size, especially for training data, which is incompatible with learning a complex model without increasing the risk of overfitting [73]. This was observed as a significant reduction in ANN classifiers performance using all available features (k=554) or the data-selected ones (k=229). Again, we speculate the reduced dimensionality and data redundancy when using physiology driven features reduced the complexity of the ANN classifiers, yielding a more generalizable model.

Finally, it is worth noting that using ANN classifiers consistently yielded higher classification accuracy when compared to LR classifiers, with or without any feature subset selection (Figure 2). However, this gain in accuracy was negligible when using the physiology-driven features ($ANN_{65} = 0.77$ vs. $LR_{65} = 0.76$ [for test set]). Given that LR classifiers are easily interpretable, our results suggest that using an LR65 classifier with physiology-driven features can yield a fully understandable decision support tool for clinical use.

5.2 Overlap in Features between Feature Selection Approaches

The secondary aim of this study was to explore whether data-driven feature selection techniques might identify ECG features indicative of ACS that were overlooked by domain-specific human experts. Table 2 mapped the 229 data-driven features against the major

components of the 12-lead ECG signal, identifying the overlap between the data-driven features and the ones selected by domain-specific expertise. More interestingly, this table summarizes the cluster of data-driven features that were overlooked by human-experts. Some of these overlooked data-driven features are contextually understandable, like ST slope, ST deviation morphology, and T wave attributes, but some other features were more challenging to classify. Upon careful annotation, we classified the overlooked data-driven features in one of these three broad categories: (1) noise attributed to existing comorbidities or patient medications (i.e., lead-specific P duration, P amplitude, and PR interval); (2) redundant information quantified by simultaneous ECG features (i.e., lead-specific Q, R, and S wave attributes that are redundant with scar size, and lead-specific QRS duration and QT interval that are redundant with principal component analysis); and (3) features that could be mechanistically linked to myocardial ischemia and can serve as plausible features of ACS (i.e., presence of fragmented QRS and lead-specific ventricular activation time).

6.0 Conclusions and Future Work[1]

6.1 Conclusions

The critical nature of clinical-decision taking results in more severe expectations from machine learning algorithms in reducing the number of false positive cases and false negative cases. Therefore, clinicians are tempted to input as much expertise as possible to "help" algorithms through the learning process by directing them towards exploring specific information-rich data. However, this approach seems to inhibit the algorithms from extensively examining all the actual available data and to narrow the scope of possibilities in terms of learnt behavior.

This prospective analysis allowed us to investigate a feature selection technique based on Cohen's d effect size, RFE and LASSO algorithms, which produced a reduced subset of ECG features. The examination of algorithms' performances using the obtained subset was accompanied with a confrontation of the results to the ones of two other versions of the data set (i.e. initial full-feature data set and physiology-driven reduced data set). Upon this comparison, we concluded that the LR classifier led by domain-specific knowledge outperformed the other classifiers and is, for this reason, the most adequate to be integrated into a clinical-decision support tools. Nevertheless, a subgroup of novel features determined thanks to data-driven approaches would give an insightful contribution in inventing new cardiac electrical biomarkers for the prompt identification of ischemia.

[1]Portions of this chapter are taken from a forthcoming paper cited, in its current status, as: **Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway,E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department,"*Journal of American Heart Association*, 2020, forthcoming.** (Available in reference [1] (submitted, under review)).

6.2 Future Work

In this analysis, the patient to feature ratio was realtively low ($\approx$ 1:1 for one of the classifiers). Fortunately, the EMPIRE study is still ongoing, the second cohort is, by this date, completed and comprise in total 1650 patients as well as a third cohort is available to test this analysis on its 923 patients. Thus, although the ECG feature selection analysis is concluded, we can apply the same algorithms on the new patients' data sets to verify the scalability of this approach and how well the features captured at this preliminary analysis are generalizable on a higher number of patients. In this same context, while the three first cohorts are ready, another cohort is in progress and would supplement our global data base with 809 more patients. Since the number of patients available is expanding, it would be interesting to try implementing deep learning techniques for acute coronary syndrome diagnosis using raw ECG data.

Furthermore, the provenance of the data set raises some concern related to a potential bias due to disparities proper to sex, race or other factors. In fact, the data collection was limited to multiple healthcare centers of one same region. An upcoming measure would be to try validating the algorithms' performance using diverse data representative of populations in remote healthcare centers.

Moreover, only ECGs collected from the paramedical staff are included in our cohorts, it would be then interesting to explore the performance of our algorithms on the general emergency department population i.e. patients presenting to the emergency department with chest pain either transported by an ambulance or walking in without calling one.

Besides, the limitations cited above added to the severe data set skew (15.3% prevalence of ACS), would have a substantial impact on the classifiers. To tackle the issue of data unbalance, we implemented artificial ACS patients' oversampling techniques which failed to generalize to an unseen testing set. Thus, future research should focus on enrolling almost equal or at least comparable proportions of negative and positive ACS patients. Once this goal reached, we should repeat the study using the new data set in order to verify the replicability of the results and get a step further towards the integration of a clinical decision support tool in electrocardiographic apparatus in emergency departments.

Finally, it would be interesting to explore other techniques for handling missing data since data imputation with the mean is a basic method that is not recommended in favor of more sophisticated procedures [62].

Bibliography

[1] Z. Bouzid, Z. Faramand, R. Gregg, S. Frisch, C. Martin-Gill, S. Saba, C. Callaway, E. Sejdic, and S. Al-Zaiti, "In search of optimal subset of ECG features to augment the diagnosis of acute coronary syndrome at the emergency department," *Journal of American Heart Association*, 2020, forthcoming.

[2] M. Y. A. Yiadom, C. W. Baugh, C. M. McWade, X. Liu, K. J. Song, B. W. Patterson, C. A. Jenkins, M. Tanski, A. M. Mills, G. Salazar *et al.*, "Performance of emergency department screening criteria for an early ECG to identify ST-segment elevation myocardial infarction," *Journal of the American Heart Association*, vol. 6, no. 3, p. e003528, 2017.

[3] E. Antman, J.-P. Bassand, W. Klein, M. Ohman, J. L. Lopez Sendon, L. Rydén, M. Simoons, and M. Tendera, "Myocardial infarction redefined—a consensus document of The Joint European Society of Cardiology/American College of Cardiology committee for the redefinition of myocardial infarction," *Journal of the American College of Cardiology*, vol. 36, no. 3, pp. 959–969, 2000.

[4] A. Mokhtari, E. Dryver, M. Söderholm, and U. Ekelund, "Diagnostic values of chest pain history, ECG, troponin and clinical gestalt in patients with chest pain and potential acute coronary syndrome assessed in the emergency department," *Springerplus*, vol. 4, no. 1, p. 219, 2015.

[5] K. Nikus, Y. Birnbaum, M. Eskola, S. Sclarovsky, Z. Zhong-qun, and O. Pahlm, "Updated electrocardiographic classification of acute coronary syndromes," *Current Cardiology Reviews*, vol. 10, no. 3, pp. 229–236, 2014.

[6] S. S. Al-Zaiti, L. Besomi, Z. Bouzid, Z. Faramand, S. O. Frisch, C. Martin-Gill, R. Gregg, S. Saba, C. Callaway, and E. Sejdić, "Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram," *Nature Communications*, 2020, in press.

[7] R. Body, G. Cook, G. Burrows, S. Carley, and P. S. Lewis, "Can emergency physicians 'rule in'and 'rule out'acute myocardial infarction with clinical judgement?" *Emergency Medicine Journal*, vol. 31, no. 11, pp. 872–876, 2014.

[8] E. P. Hess, D. Agarwal, S. Chandra, M. H. Murad, P. J. Erwin, J. E. Hollander, V. M. Montori, and I. G. Stiell, "Diagnostic accuracy of the TIMI risk score in patients with chest pain in the emergency department: a meta-analysis," *CMAJ*, vol. 182, no. 10, pp. 1039–1044, 2010.

[9] E. P. Hess, R. J. Brison, J. J. Perry, L. A. Calder, V. Thiruganasambandamoorthy, D. Agarwal, A. T. Sadosty, M. L. Silvilotti, A. S. Jaffe, V. M. Montori *et al.*, "Development of a clinical prediction rule for 30-day cardiac events in emergency department patients with chest pain and possible acute coronary syndrome," *Annals of Emergency Medicine*, vol. 59, no. 2, pp. 115–125, 2012.

[10] S. S. Al-Zaiti, V. Shusterman, and M. G. Carey, "Novel technical solutions for wireless ECG transmission & analysis in the age of the internet cloud," *Journal of Electrocardiology*, vol. 46, no. 6, pp. 540–545, 2013.

[11] Y. Birnbaum, J. M. Wilson, M. Fiol, A. B. de Luna, M. Eskola, and K. Nikus, "ECG diagnosis and classification of acute coronary syndromes," *Annals of Noninvasive Electrocardiology*, vol. 19, no. 1, pp. 4–14, 2014.

[12] T. Quinn, S. Johnsen, C. P. Gale, H. Snooks, S. McLean, M. Woollard, and C. Weston, "Effects of prehospital 12-lead ECG on processes of care and mortality in acute coronary syndrome: a linked cohort study from the Myocardial Ischaemia National Audit Project," *Heart*, vol. 100, no. 12, pp. 944–950, 2014.

[13] G. S. Wagner, P. Macfarlane, H. Wellens, M. Josephson, A. Gorgels, D. M. Mirvis, O. Pahlm, B. Surawicz, P. Kligfield, R. Childers, and L. S. Gettes, "AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram," *Journal of the American College of Cardiology*, vol. 53, no. 11, pp. 1003–1011, 2009.

[14] S. S. Al-Zaiti, C. Martin-Gill, E. Sejdić, M. Alrawashdeh, and C. Callaway, "Rationale, development, and implementation of the electrocardiographic methods for the prehospital identification of non-ST elevation myocardial infarction events (EMPIRE)," *Journal of Electrocardiology*, vol. 48, no. 6, pp. 921–926, 2015.

[15] R. L. Lux, "Non-ST-segment elevation myocardial infarction: A novel and robust approach for early detection of patients at risk," *Journal of the American Heart Association*, vol. 4, no. 7, p. e002279, 2015.

[16] K. Thygesen, J. S. Alpert, A. S. Jaffe, B. R. Chaitman, J. J. Bax, D. A. Morrow, H. D. White, H. Mickley, F. Crea, F. Van de Werf *et al.*, "Fourth universal definition

of myocardial infarction (2018)," *European Heart Journal*, vol. 40, no. 3, pp. 237–269, 2019.

[17] P. J. Leisy, R. R. Coeytaux, G. S. Wagner, E. H. Chung, A. J. McBroom, C. L. Green, J. W. Williams Jr, and G. D. Sanders, "ECG-based signal analysis technologies for evaluating patients with acute coronary syndrome: A systematic review," *Journal of Electrocardiology*, vol. 46, no. 2, pp. 92–97, 2013.

[18] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[19] R. Body, "Acute coronary syndromes diagnosis, version 2.0: Tomorrow's approach to diagnosing acute coronary syndromes?" *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 94–99, 2018.

[20] K. Thygesen, J. S. Alpert, A. S. Jaffe, B. R. Chaitman, J. J. Bax, D. A. Morrow, H. D. White *et al.*, "Fourth universal definition of myocardial infarction (2018)," *Journal of the American College of Cardiology*, vol. 72, no. 18, pp. 2231–2264, 2018.

[21] J. J. Bax, H. Baumgartner, C. Ceconi, V. Dean, R. Fagard, C. Funck-Brentano, D. Hasdai, A. Hoes, P. Kirchhof, J. Knuuti *et al.*, "Third universal definition of myocardial infarction," *Journal of the American College of Cardiology*, vol. 60, no. 16, pp. 1581–1598, 2012.

[22] J. S. Smith, M. K. Cahalan, D. J. Benefiel, B. Byrd, F. Lurz, W. Shapiro, M. Roizen, A. Bouchard, and N. Schiller, "Intraoperative detection of myocardial ischemia in high-risk patients: electrocardiography versus two-dimensional transesophageal echocardiography." *Circulation*, vol. 72, no. 5, pp. 1015–1021, 1985.

[23] C. Berry, A. Zalewski, R. Kovach, M. Savage, and S. Goldberg, "Surface electrocardiogram in the detection of transmural myocardial ischemia during coronary artery occlusion," *American Journal of Cardiology*, vol. 63, no. 1, pp. 21–26, 1989.

[24] P. M. Rautaharju, S. H. Zhou, E. W. Hancock, B. M. Hor, D. Q. Feild, J. M. Lindauer, G. S. Wagner, O. Pahlm, C. L. Feldman *et al.*, "Comparability of 12-lead ECGs derived from EASI leads with standard 12-lead ECGs in the classification of acute myocardial ischemia and old myocardial infarction," *Journal of Electrocardiology*, vol. 35, no. 4, pp. 35–39, 2002.

[25] S. S. Al-Zaiti, M. Alrawashdeh, C. Martin-Gill, C. Callaway, D. Mortara, and J. Nemec, "Evaluation of beat-to-beat ventricular repolarization lability from standard 12-lead ECG during acute myocardial ischemia," *Journal of Electrocardiology*, vol. 50, no. 6, pp. 717–724, November–December 2017.

[26] A. Kölliker and H. Müller, "Nachweis der negativen Schwankung des Muskelstroms am naturlich sich contrahirenden Muskel," *Verhandlungen Physikalisch-Medizinische Gesellschaft*, vol. 6, pp. 528–533, 1856.

[27] M. Rowbottom and C. Susskind, *Electricity and Medicine: History of their Interaction.* San Francisco, CA, USA: San Francisco Press, 1984.

[28] X.-L. Yang, G.-Z. Liu, Y.-H. Tong, H. Yan, Z. Xu, Q. Chen, X. Liu, H.-H. Zhang, H.-B. Wang, and S.-H. Tan, "The history, hotspots, and trends of electrocardiogram," *Journal of Geriatric Cardiology*, vol. 12, no. 4, p. 448–456, 2015.

[29] D. E. Becker, "Fundamentals of electrocardiography interpretation," *Anesthesia Progress*, vol. 53, no. 2, pp. 53–64, 2006.

[30] R. J. Noble, J. S. Hillis, and D. A. Rothbaum, "Electrocardiography," in *Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition.*, H. K. Walker, W. D. Hall, and J. W. Hurst, Eds. Boston, MA: Butterworths, 1990, ch. 33.

[31] T. B. Garcia, *12-Lead ECG: The Art of Interpretation, 2nd edition.* Burlington, MA: Jones and Bartlett Learning, 2013.

[32] C. P. Jayapandian, "Cloudwave: A Cloud Computing Framework for Multimodal Electrophysiological Big Data," Ph.D. dissertation, Case Western Reserve University, 2014.

[33] J. Song, H. Yan, Z. Xu, X. Yu, and R. Zhu, "Myocardial ischemia analysis based on electrocardiogram QRS complex," *Australasian Physical and Engineering Sciences in Medicine*, vol. 34, no. 4, pp. 515–521, 2011.

[34] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959.

[35] T. Mitchell, *Machine Learning.* New York, NY: McGraw Hill, Inc., 1997.

[36] D. Wang, "Unsupervised learning: Foundations of neural computation," *AI Magazine*, vol. 22, no. 2, pp. 101–102, 2001.

[37] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. MIT press, 2018.

[38] D. O. Hebb, *The Organization of Behavior*. New York, NY: Wiley, 1949.

[39] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.

[40] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, no. 1, pp. 1–11, 1968.

[41] A. Luntz and V. Brailovsky, "On estimation of characters obtained in statistical procedure of recognition," *Technicheskaya Kibernetica*, vol. 3, 1969.

[42] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognition*, vol. 48, no. 9, pp. 2839–2846, September 2015.

[43] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[44] X. Zeng and T. R. Martinez, "Distribution-balanced stratified cross-validation for accuracy estimation," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 12, no. 1, pp. 1–12, 2000.

[45] P. McCullagh, *Generalized linear models*. Routledge, 2018.

[46] J. McGonagle, G. Shaikouski, C. Williams, A. Hsu, and J. Khim, "Backpropagation," *Brilliant.org*, Available: https://brilliant.org/wiki/backpropagation/. Accessed: 4-May-2020.

[47] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press, 2007.

[48] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv preprint arXiv:1811.03378*, 2018.

[49] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, January 2015.

[50] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., 1995.

[51] N. Sharma and K. Saroha, "Study of dimension reduction methodologies in data mining," in *International Conference on Computing, Communication and Automation, ICCCA 2015, Noida, India, May 15-16, 2015*. IEEE, 2015, pp. 133–137.

[52] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, 2019.

[53] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, p. 1, 2010.

[54] R. Liu and D. F. Gillies, "Overfitting in linear feature extraction for classification of high-dimensional image data," *Pattern Recognition*, vol. 53, pp. 73–86, May 2016.

[55] D. M. Hawkins, "The problem of overfitting," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[56] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[57] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[58] K. Hajian-Tilaki, "Sample size estimation in diagnostic test studies of biomedical informatics," *Journal of Biomedical Informatics*, vol. 48, pp. 193–204, April 2014.

[59] N. M. F. Buderer, "Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity," *Academic Emergency Medicine*, vol. 3, no. 9, pp. 895–900, 1996.

[60] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019, vol. 793.

[61] J. W. Graham, "Missing Data Theory," in *Missing Data. Statistics for Social and Behavioral Sciences*. New York, NY: Springer, 2012, pp. 3–46.

[62] ——, "Missing data analysis: Making it work in the real world," *Annual Review of Psychology*, vol. 60, pp. 549–576, 2009.

[63] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado, "The use of receiver operating characteristic curves in biomedical informatics," *Journal of Biomedical Informatics*, vol. 38, no. 5, pp. 404–415, 2005.

[64] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine," *Clinical Chemistry*, vol. 39, no. 4, pp. 561–577, 04 1993.

[65] Z. Zhang, "Statistical considerations for evaluating prognostic biomarkers: Choosing optimal threshold," in *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*. Springer, 2016, pp. 15–20.

[66] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[67] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the Youden Index and its associated cutoff point," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 47, no. 4, pp. 458–472, 2005.

[68] R. Froud and G. Abel, "Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: The forgotten lesson of Pythagoras. Theoretical considerations and an example application of change in health status," *PLoS One*, vol. 9, no. 12, p. e114468, 2014.

[69] J. L. Forberg, M. Green, J. Björk, M. Ohlsson, L. Edenbrandt, H. Öhlin, and U. Ekelund, "In search of the best method to predict acute coronary syndrome using only the electrocardiogram from the emergency department," *Journal of Electrocardiology*, vol. 42, no. 1, pp. 58–63, 2009.

[70] M. Green, J. Björk, J. Forberg, U. Ekelund, L. Edenbrandt, and M. Ohlsson, "Comparison between neural networks and multiple logistic regression to predict acute coro-

nary syndrome in the emergency room," *Artificial Intelligence in Medicine*, vol. 38, no. 3, pp. 305–318, 2006.

[71] C.-C. Wu, W.-D. Hsu, M. M. Islam, T. N. Poly, H.-C. Yang, P.-A. A. Nguyen, Y.-C. Wang, and Y.-C. J. Li, "An artificial intelligence approach to early predict non-ST-elevation myocardial infarction patients with chest pain," *Computer Methods and Programs in Biomedicine*, vol. 173, pp. 109–117, May 2019.

[72] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated Receiver Operating Characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, September 1988.

[73] P. D. Myers, B. M. Scirica, and C. M. Stultz, "Machine learning improves risk stratification after acute coronary syndrome," *Scientific Reports*, vol. 7, no. 12692, pp. 1–12, 2017.