ECG Feature Importance Rankings: Cardiologists vs. Algorithms

Temesgen Mehari[1,2], Ashish Sundar[3], Alen Bosnjakovic[4], Peter Harris[3], Steven E. Williams[5], Axel Loewe[6], Olaf Doessel[6], Claudia Nagel[6], Nils Strodthoff[7], Philip J. Aston[3,8]

***Abstract*—Feature importance methods promise to provide a ranking of features according to importance for a given classification task. A wide range of methods exist but their rankings often disagree and they are inherently difficult to evaluate due to a lack of ground truth beyond synthetic datasets. In this work, we put feature importance methods to the test on real-world data in the domain of cardiology, where we try to distinguish three specific pathologies from healthy subjects based on ECG features comparing to features used in cardiologists' decision rules as ground truth. We found that the SHAP and LIME methods and Chi-squared test all worked well together with the native Random forest and Logistic regression feature rankings. Some methods gave inconsistent results, which included the Maximum Relevance Minimum Redundancy and Neighbourhood Component Analysis methods. The permutation-based methods generally performed quite poorly. A surprising result was found in the case of left bundle branch block, where T-wave morphology features were consistently identified as being important for diagnosis, but are not used by clinicians.**

***Index Terms*—Electrocardiogram, feature importance ranking, cardiologist, atrioventricular block, right branch bundle block, left branch bundle block.**

I. Introduction

A trained cardiologist can diagnose over 150 different conditions from a 12-lead electrocardiogram (ECG) [1]. Such diagnoses are made on the basis of a multitude of ECG features which consist mainly of time intervals between certain fiducial points on the ECG, amplitudes of prominent features or morphology of ECG segments. For each pathology, the relevant criteria for specific features are well documented [1], [2], although there may be minor differences between one reference source and another.

On the other hand, there are numerous algorithms available for determining a ranking of features by importance for a given classification task [3]. However, if several algorithms are used, then it is often found that they give significantly different feature importance rankings and it is not apparent which ranking is best or whether one particular ranking is better than another. Therefore, we did a comparison of feature importance rankings generated by a number of different algorithms with the corresponding features that a cardiologist uses for diagnosis. This has the advantage of having a set of important features which has been gleaned from clinical experience over many years for the diagnosis of each condition which can be compared with the feature rankings of the algorithms.

Another possibility with this study is that the feature importance algorithms could identify features that are important for the diagnosis of a condition which are not normally considered to be important by cardiologists.

We have chosen three pathologies to study, namely first degree atrioventricular block (1st degree AV block), complete right bundle branch block (RBBB) and complete left bundle branch block (LBBB). A diagnosis of these conditions by cardiologists involves 1, 7 and 14 features respectively and so are progressively more complex, starting with the simplest possible case. In addition, all three pathologies are commonly encountered in the general population and are well-represented in the PTB-XL dataset underlying our study.

For this study, we restrict attention to the simplest case of a binary classification that seeks to distinguish healthy subjects vs. a specific pathology. Of course in practice, a cardiologist has to identify a condition (or multiple conditions) out of many possible conditions, which is a much more complicated task. On the other hand, it is quite conceivable that a simple binary classification of healthy vs. a specific pathology could be successfully achieved by using only a reduced subset of the complete list of diagnostic conditions. However, we consider it appropriate to study the simplest case first. A study of multiclass feature importance algorithms with all four of the above classes has been undertaken as a separate study [4].

We are considering the features used by cardiologists for diagnosis to be the gold standard against which we compare various algorithms. However, it should be noted that different sources for ECG diagnosis often give slightly different conditions for the diagnosis of a specific pathology. This may be because textbooks give sufficient conditions for diagnosis, rather than an exhaustive list of all changes associated with a pathology. We have used *EKG-Kurs für Isabel* [5] as it gives simple, itemised conditions for each pathology. More comprehensive texts are available but we chose this one based on its simplicity and clarity.

An alternative approach to identifying important parts of the ECG signal for diagnosis of a particular condition is to use explainable AI (XAI) methods applied to models operating

This project (18HLT07 MedalCare) has received funding from the EMPIR programme co-financed by the Participating States and from the European Union's Horizon 2020 research and innovation programme.

[1] Heinrich Hertz Institute, Berlin, Germany
[2] Physikalisch-Technische Bundesanstalt, Berlin, Germany
[3] National Physical Laboratory, Teddington, UK
[4] Institute of Metrology of Bosnia and Herzegovina, Sarajevo, Bosnia and Herzegovina
[5] University of Edinburgh, Edinburgh, UK
[6] Karlsruhe Institute of Technology, Karlsruhe, Germany
[7] University of Oldenburg, Oldenburg, Germany
[8] University of Surrey, Guildford, UK

on the raw signal [6], [7]. We present a detailed comparison between both approaches in the Discussion Section.

We believe that while feature importance methods are well-established, their application and systematic evaluation in cardiology has not been covered in the literature. Our work addresses this gap by:

- using feature importance methods to rank ECG features based on their ability to discriminate specific cardiac pathologies from healthy patients. By juxtaposing these with cardiologists' clinical decision rules, we shed light on the alignment (or lack thereof) between computational methods and real-world clinical practice.
- providing a comprehensive comparison of these methods, which is rare in the existing literature, especially in the field of cardiology. Such a comprehensive comparison provides cardiology practitioners with a guide to help them determine which methods are most reliable and applicable in specific clinical settings.
- uncovering new insights from feature importances. Notably, our study reveals a surprising finding: Certain features of T-wave morphology consistently emerge as critical for the diagnosis of LBBB, in contrast to prevailing clinical practice, which relies predominantly on QRS-complex-related features for this diagnosis.

It is worth stressing that this work does not line up in the approaches which try to enhance decision support systems with XAI side information. Rather than directly advocating for the integration of XAI into diagnostic tools, this study serves as a critical evaluation of feature importance methods, particularly for tabular classifiers and underscores the potential of XAI to be used in knowledge discovery.

II. Materials and Methods

A. ECG Signals

The ECG signals that were used for this study were taken from the PTB-XL dataset [8], [9], which is publicly available on PhysioNet [10]. In particular, for each of the three pathologies considered (1st degree AV block, RBBB, LBBB), we extracted all the records that were labelled with only the specific pathology.

B. ECG Features

For extracting features from an ECG, we used the University of Glasgow 12-lead ECG analysis algorithm which has been developed over many years by a team at the University of Glasgow [11]. This software can derive more than 772 global and lead-dependent ECG features from a 10-second 12-lead ECG signal. (All the features derived by the Glasgow software for the PTB-XL dataset are available in the PTB-XL+ feature dataset [12], [10].) From this large collection of features, we selected 117 which a cardiologist would typically assess when considering a diagnosis that are given in Appendix A. The selection of features may be subject to discussion, and some might advocate for the inclusion of different features. However, it is important to note that there is not a universally agreed upon set of features. These features were derived for all of the ECG records in each of the pathology classes. The small number of records that contained missing values due to issues with feature extraction were deleted to obtain a final dataset without missing values. Features were also drawn from an equal number of healthy patients' records, chosen at random. If any records had missing values, they were replaced by other records sampled at random. With this approach, a balanced dataset containing no missing values was created for each pathology.

Each feature was scaled to have a mean of zero and a variance of one to give a standardized dataset, which was required for certain algorithms (Logistic regression) or is known to be beneficial for others (Deep networks).

The final datasets contained a total of 1,592 records for 1st degree AV block, 1,074 records for RBBB and 1,072 records for LBBB, with half being for healthy subjects and half for the specific pathology in each case.

C. Pathologies

The ECG is the difference in electrical potential measurable between two different electrodes attached to the body surface and captures the electrical activity due to de- and repolarization of cardiomyocytes in the heart. In the healthy case, electrical activity is spontaneously initiated in the pacemaker cells at the sinoatrial node in the right atrium. After spreading throughout the atrial myocardial tissue and causing the P wave in the ECG, the excitation is delayed at the atrioventricular node. The electrical activation is then conducted via the bundle of His, which branches into a right bundle as well as an anterior and a posterior left bundle before it reaches the Purkinje fibers. These activate the ventricular myocardium from the apex to the base and lead to the QRS complex in the ECG. Finally, the T wave in the ECG arises due to repolarization of the ventricular myocytes.

We now consider each of our chosen pathologies in detail.

1) Atrioventricular Block: In patients with atrioventricular block, the excitation conduction between atria and ventricles is impaired. In first degree AV block, which is studied in this work, the conduction is markedly delayed and leads to PR intervals $>$200 ms in the ECG. However, all atrial impulses are still transferred to the ventricles and every P wave is followed by a QRS complex as opposed to second or third degree AV-block that is associated with skipped beats or independent excitation of atria and ventricles respectively [5]. Thus, there is only one feature which is used for the diagnosis of a 1st degree AV block:

- PR interval

We checked for other features that correlate (with absolute Pearson correlation coefficient ≥ 0.7) with the PR interval, as such features may be expected to occur high up the ranking. However, there were none and so this is the simplest possible case.

2) Right Bundle Branch Block: Complete right bundle branch block is characterized by a marked delay or block in conduction in the right bundle branch. In this case, the right ventricles are activated via impulses conducted through the left bundle branches reaching the right ventricle through

TABLE I
FEATURES THAT CORRELATE WITH THE IMPORTANT FEATURES FOR RBBB, BUT NOT INCLUDING OTHER IMPORTANT FEATURES, TOGETHER WITH THEIR CORRELATION COEFFICIENTS.

Feature	Correlating Features (correlation coefficient)
R amplitude, lead V1	–
R' amplitude, lead V1	Peak-to-peak amplitude, lead V1 (0.79)
S amplitude, lead I	–
S amplitude, lead aVL	–
S amplitude, lead V1	–
S amplitude, lead V6	R' amplitude, lead V5 (-0.71) S amplitude, lead V5 (0.82)
QRS duration	–

the ventricular myocardial tissue. As this takes longer than the physiological activation through the three fascicles, this reflects in a widened QRS complex of $>120\,\text{ms}$ in the ECG. Furthermore, a terminal R' peak is visible in lead V1 and a notched S wave occurs in leads I, aVL and V6 [5]. Thus, the 7 features that are relevant for diagnosis of right bundle branch block are:

- QRS duration
- R amplitude in lead V1
- R' amplitude in lead V1
- S amplitude in leads I, aVL, V1 and V6

We call these 7 features the important features for RBBB. We checked for features that correlate (with absolute Pearson correlation coefficient $\geq$ 0.7) with one of these 7 features. There were 3 such features, not including the important features above, which are given in Table I.

3) Left Bundle Branch Block: Analogously to right bundle branch block described above, complete left bundle branch block describes the condition of a blockage in the electrical conduction in the left bundle branch. As the left bundle branches into an anterior and a posterior fascicle, the term complete left bundle branch block refers to a conduction block before the bifurcation. In the ECG, the delayed activation of the left ventricle reflects in a widened QRS complex of $>120\,\text{ms}$, deep Q waves in lead V1 and a notched or monophasic QRS morphology in the lateral leads I, aVL, V5 and V6 [5]. Thus, there are 14 features that are involved in the diagnosis of left bundle branch block:

- QRS duration
- Q amplitude in lead V1
- R amplitude in leads I, aVL, V5 and V6
- R' amplitude in leads I, aVL, V5 and V6
- S amplitude in leads I, aVL, V5 and V6

We call these 14 features the important features for LBBB. We checked for features that correlate (with absolute Pearson correlation coefficient $\geq$ 0.7) with one of these 14 features, excluding the important features listed above. Table II lists the 28 features identified through this analysis.

D. Feature Importance Algorithms

We can broadly categorize the feature importance algorithms investigated in this work as model-dependent and model-independent methods.

TABLE II
FEATURES THAT CORRELATE WITH THE IMPORTANT FEATURES FOR LBBB, BUT NOT INCLUDING OTHER IMPORTANT FEATURES, TOGETHER WITH THEIR CORRELATION COEFFICIENTS.

Feature	Correlating Features (correlation coefficient)
QRS duration	Q amplitude, lead V4 (-0.71) S amplitude, lead V3 (-0.71) T+ amplitude, lead V1 (0.75) ST slope, lead I (-0.73) ST slope, lead V1 (0.77) ST slope, lead V6 (-0.70) ST duration (-0.74) T morphology, lead I (-0.82) T morphology, lead aVR (0.77) T morphology, lead V6 (-0.79)
Q amplitude, lead V1	Peak-to-peak amplitude, lead V1 (-0.90) T+ amplitude, lead V1 (-0.79)
R amplitude, lead I	Peak-to-peak amplitude, lead I (0.94)
R amplitude, lead aVL	Peak-to-peak amplitude, lead I (0.73) Peak-to-peak amplitude, lead aVL (0.90) Q amplitude, lead III (-0.87) Q amplitude, lead aVF (-0.79) S amplitude, lead III (-0.81) QRS frontal axis (-0.72)
R amplitude, lead V5	Peak-to-peak amplitude, lead V5 (0.86) Peak-to-peak amplitude, lead V6 (0.78) R amplitude, lead V4 (0.77)
R amplitude, lead V6	Peak-to-peak amplitude, lead V5 (0.72) Peak-to-peak amplitude, lead V6 (0.95)
R' amplitude, lead I	Peak-to-peak amplitude, lead aVL (0.75)
R' amplitude, lead aVL	Peak-to-peak amplitude, lead aVL (0.72) S amplitude, lead II (-0.70) S amplitude, lead III (-0.78) S amplitude, lead aVF (-0.78)
R' amplitude, lead V5	–
R' amplitude, lead V6	–
S amplitude, lead I	R' amplitude, lead aVR (-0.71) T+ amplitude, lead aVR (-0.77)
S amplitude, lead aVL	R amplitude, lead III (-0.74) R' amplitude, lead II (-0.74) R' amplitude, lead III (-0.72)
S amplitude, lead V5	–
S amplitude, lead V6	R' amplitude, lead V1 (-0.81)

TABLE III
HYPERPARAMETERS FOR THE MACHINE LEARNING MODELS USED.

Model	Hyperparameters
Random Forests	number of trees = 100 criterion = 'Gini impurity'
Boosted Decision Trees (XGB)	loss function = 'binary logistic' learning rate = 0.01 early stopping rounds = 20 number of boosting iterations = 5000

Logistic Regression	loss = 'l_2 norm'
	tol = 0.0001
	max_iter = 100

Deep Neural Networks	2 hidden fully-connected layers with dim = 256 hidden activations ='relu' output activation = 'sigmoid' optimizer = 'adam' loss = 'binary cross entropy'

1) Model-dependent feature importance methods:

- **Random forests, Boosted decision trees, Logistic regression and Deep neural networks with permutation/SHAP/LIME feature importance.** In terms of models, we consider Random forests, Boosted decision trees, Logistic regression and Deep neural networks. The hyperparameters used are summarized in Table III.

They correspond to the default parameters, except for the deep neural network, where we started with a very deep network (10 layers) and removed layers as long as there was no noticeable drop in performance. The training data consisted of records from the PTB-XL stratified folds 1–9 and the test data were records drawn from fold 10 [8]. These models were then combined with established attribution methods LIME [13], SHAP [14] and permutation feature importance [15]. LIME involves training an interpretable, local surrogate model to approximate the model behaviour near the sample of interest. SHAP is an efficient implementation of the game-theoretic Shapley value approach. LIME and SHAP are local attribution methods which return attribution scores per sample, and we therefore ranked the features on the mean of the absolute attribution values across the test set. As a third class of feature importance algorithms, we considered permutation feature importance, a global attribution method which quantifies feature importance via the decrease in model performance upon replacing a feature column of interest by a permuted copy of itself.

- **Random forests.** For a random forest model, the importance of features can be determined by how much they decrease Gini impurity when averaged over all the trees in the forest. It is known that these feature importance values can be misleading for high cardinality features. However, permutation feature importance (see below) can mitigate this to some extent [16].
- **Logistic regression.** The importance of features in a logistic regression model can be determined by the exponential of the weight associated with each feature [3].
- **Gaussian processes.** In Gaussian Process binary classification, the probability of class membership conditioned on an observed feature vector x is modelled as $\sigma(f)$ where σ is a sigmoid function, such as the logistic function, and a Gaussian Process model GP$(0, k(x, x'))$ is used as a prior distribution for the latent variable f [17]. Using a squared exponential covariance kernel for $k(x, x')$ with diagonal covariance matrix, each feature x_i is associated with its own length-scale parameter l_i. A small value for l_i implies the feature varies over short-length scales and so is important for the classification. Consequently, sorting the length-scale parameters provides a ranking of the features.

  In our implementations, we employed Gaussian Process regression, designating $+1$ for the positive class and -1 for the negative class, and using the sigmoid function to infer hard predictions. For inference, specifically for approximating the integral in the posterior, we adopted Laplace's method. We refer to the code repository for further implementation details. [18]

2) Model-independent feature importance methods: In addition to model-dependent methods, we also include methods that solely rely on the data distribution without making use of a trained predictor on the dataset. In the feature selection literature [19], [20], [21], these methods are often referred to as filter methods. More specifically, we consider the following methods:

- **Chi-square test.** The Chi-square test is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis. Each feature is tested individually for independence of the response. A small p-value is associated with a feature that has dependence on the response, and so is important. Thus, features are ranked by $-\log(p_i)$, where i is the index of the features [22].
- **Maximum Relevance - Minimum Redundancy (MRMR).** The MRMR method reduces redundant features while keeping the relevant features for the model, where redundancy and relevance are quantified in terms of mutual information. It is known that many essential features are correlated and redundant and so the MRMR method selects features taking into account the relevance for predicting the outcome variable and the redundancy within the selected features [23], [24].
- **Neighbourhood Component Analysis (NCA).** The NCA method selects features by maximizing the prediction accuracy of classification algorithms. The concept of this method is similar to the k-nearest neighbours classification method, only in the NCA method, the reference point is selected randomly not to be the nearest neighbour for the new point [25].
- **ReliefF.** ReliefF calculates a feature score for each feature depending on feature value differences for neighbours which have the same or a different class, which can then be used to rank the features. The ReliefF method estimates the attribute qualities based on how well they can distinguish between instances near them. This method was initially designed to apply to binary classification problems with discrete or numerical features [26].
- **Modified ROC AUC.** The receiver operating characteristic (ROC) curve consists of a plot of the false positive rate against the true positive rate as a threshold is moved across the distributions for the two classes. The area under the curve (ROC AUC) is a standardised measure of the degree of separation of the two distributions and varies from 0.5 (no discrimination) to 1 (perfect discrimination) [27].

  We note that the positive class has to be specified and that if this is changed from one class to the other, the ROC AUC values range from 0.5 (no discrimination) to 0 (perfect discrimination). Thus, we define $\text{Modified ROC AUC} = \max(\text{ROC AUC}, 1\text{- ROC AUC})$. This ensures that all values are in the range of 0.5 to 1.

  A feature ranking can be generated for a binary classification problem by generating a distribution for each of the two classes for each feature individually and then finding the Modified ROC AUC for all of these distributions. The features are then ranked by their Modified ROC AUC value from highest to lowest, which ranks the features according to their ability to discriminate the two classes individually.

  We also note that the ROC AUC values aid with interpretation of the features, as $\mathrm{ROC\ AUC} > 0.5$ implies that the feature increases due to the pathology whereas $\mathrm{ROC\ AUC} < 0.5$ means that the feature decreases due to the pathology.

To ensure transparency and reproducibility, the implementations of all methods and models, that were used in this study, have been made publicly accessible at [18].

E. Scoring algorithm

When comparing a feature ranking generated by one of the algorithms with the important feature set for diagnosis of a specific pathology, we define a score that enables a simple comparison between different methods, assuming that all features in the set of important features for diagnosis have equal importance. As a first step, we choose a value of n, which is the number of top features in each ranking that will be considered. We then take a weighted average based on the ranking of each of the top n features that is contained in the important set, where the first feature has a weighting of n, the second a weighting of $n - 1$ and so on, so that the n^{th} feature has a weighting of 1. This weighted average is then normalised to give a score between 0 and 100 (by dividing by $n(n + 1)/2/100$), which we round to the nearest integer.

With this scoring system, an important feature in position 1 of the ranking contributes $200/(n+1)$ to the score, whereas an important feature in position n only contributes $200/(n(n + 1))$ to the score. For example, taking $n = 5$ and assuming that a ranking has the first, second and fourth features in the important set gives a score of $(5 + 4 + 2)/15 \times 100 \approx 73$.

We also consider the ranking of features that are least able to discriminate between the two classes, which can be defined by the features with the lowest modified ROC AUC values. In particular, we consider the two features with the lowest modified ROC AUC values which, for the three pathologies, are as follows:

- 1^{st} degree AV block: S amplitude, lead I (modified ROC AUC=0.5006); S amplitude, lead V2 (modified ROC AUC=0.5006)
- RBBB: R’ amplitude, lead V6 (modified ROC AUC=0.5028); R’ amplitude, lead I (modified ROC AUC=0.5037)
- LBBB: R amplitude, lead I (modified ROC AUC=0.5002); R’ amplitude, lead V6 (modified ROC AUC=0.5009)

We refer to these as the *non-discriminating features*.

III. Results

We consider results of the feature importance ranking algorithms applied to the feature table for each pathology in turn. The model-dependent methods first require training of a machine learning model for the binary classification problem. The accuracy of the five machine learning models for each pathology on the test data are shown in Table IV. Clearly, these are all very high. For this reason, we refrain from further hyperparameter tuning.

TABLE IV
The accuracy of the machine learning models on the test data for each distinguishing each pathology from an equally sized set of normal samples.

Model	**1^{st} degree AV block**	**RBBB**	**LBBB**
Random forest	95.6%	99.1%	100%
XGB	96.8%	98.1%	100%
Logistic regression	95.6%	100%	100%
Deep networks	94.3%	100%	100%
Gaussian processes	97.8%	100%	100%

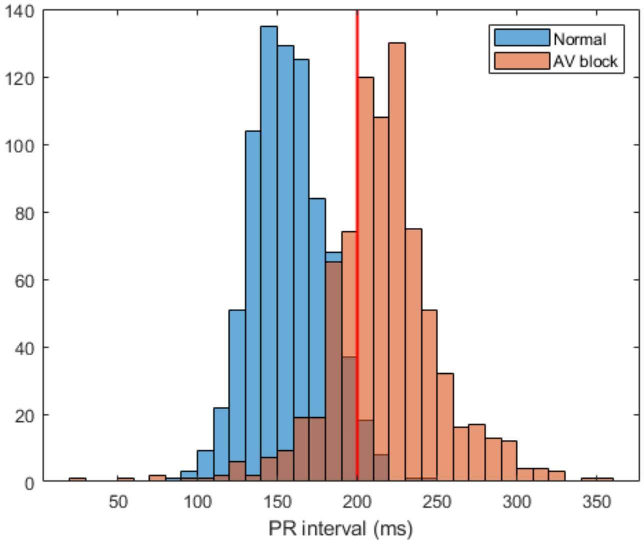

Fig. 1. Histogram of the PR interval for the records labelled as healthy and 1^{st} degree AV block. The red line is at 200 ms, which is the threshold for diagnosis of 1^{st} degree AV block.

A. Atrioventricular Block

First degree AV block is defined by the PR interval being greater than 200 ms [1]. Thus, there is a single important parameter for diagnosis in this case, namely the PR interval. We therefore expect this feature to occur high up in the rankings.

For the data we are using, the distributions for the PR interval for the records labelled as Normal and 1^{st} degree AV block are shown in Fig. 1. Clearly, not all of the 1^{st} degree AV block records satisfy the diagnostic criterion of exceeding 200 ms. In fact, the PR interval for 236 out of 796 records labelled as 1^{st} degree AV block does not exceed 200 ms, with the smallest value being 26 ms (which is non-physiological). Conversely, there are 23 out of 796 records labelled as Normal that have PR interval exceeding 200 ms, with the largest value being 242 ms. Presumably in both cases this is because the Glasgow algorithm identifies the PR interval as shorter/longer than that identified by the cardiologists who labelled the signals.

As an aside, if we tried to classify 1^{st} degree AV block using the Glasgow computed PR intervals, then the Modified ROC AUC for this classification is 0.9384 and the optimal threshold for diagnosis is 184 ms, which is considerably lower than the conventional 200 ms threshold. With this threshold,

TABLE V
RANKING OF THE PR INTERVAL AND THE NON-DISCRIMINATING FEATURES WHEN CONSIDERING NORMAL AND 1^{st} DEGREE AV BLOCK SIGNALS. RESULTS FOR MODEL-DEPENDENT METHODS ARE GIVEN IN THE UPPER PART OF THE TABLE AND RESULTS FOR MODEL-INDEPENDENT METHODS ARE GIVEN IN THE LOWER PART. WE PRESENT THE MEAN AND STANDARD DEVIATIONS AS INTEGERS, CALCULATED OVER 5 RUNS, TO ACCOUNT FOR THE INHERENT RANDOMNESS OF THE METHODS.

Method	Ranking of the PR interval	Ranking of the non-discriminating features
Random Forest (permutation)	**1(0)**	36(37), 13(2)
Random Forest (SHAP)	**1(0)**	73(6), 19(5)
Random Forest (LIME)	**1(0)**	54(23), 51(22)
Random Forest	**1(0)**	68(9), 23(10)
XGB (permutation)	**1(0)**	18(2), 24(1)
XGB (SHAP)	**1(0)**	67(0), 32(0)
XGB (LIME)	**1(0)**	56(18), 70(17)
Logistic Regression (permutation)	**1(0)**	80(19), 78(16)
Logistic Regression (SHAP)	**1(0)**	65(1), 82(1)
Logistic Regression (LIME)	**1(0)**	54(20), 60(10)
Logistic Regression	**1(0)**	76(0), 98(0)
Deep networks (permutation)	**1(0)**	63(24), 87(21)
Deep networks (SHAP)	**1(0)**	84(16), 75(15)
Deep networks (LIME)	**1(0)**	46(20), 68(13)
Gaussian processes	**1(0)**	84(0), 93(0)
Chi square test	**1(0)**	99(0), 103(0)
MRMR	**1(0)**	60(1), 87(1)
NCA	**1(0)**	15(2), 61(4)
Relieff	**1(0)**	48(0), 58((0))
Modified ROC AUC	**1(0)**	116(0), 117(0)

the accuracy of the classification is 88.13%. Presumably this reduced threshold is a result of the difference between the PR interval lengths determined by the cardiologists and the Glasgow software.

The ranking of the PR interval by each algorithm is shown in Table V. These results show that all the algorithms we considered ranked the PR interval as the most important feature.

We also considered the ranking for each method of the non-discriminating features, which both have Modified ROC AUC values very close to 0.5. These are shown in the final column of Table V. We note that they are by definition the last two features in the Modified ROC AUC ranking. We observe that some methods exhibit relatively high rankings for non-discriminating features. For instance, the second non-discriminating feature is ranked 13th for Random Forest (Permutation), 19th for Random Forest (SHAP), and 23th for Random Forest. In contrast, the Logistic Regression, Gaussian Processes and Chi square test all rank the same feature above 90.

We then found the top 5 features for each of the methods to see if there is any commonality between them. The frequency of features in the top 5 is shown in Table VI which, as expected, includes the PR interval as the most common. When averaging the rankings across all runs, two methods had their top 5 features matching those in Table VI, which were Random forest and Random forest (SHAP), while Random forest (LIME), XGB (SHAP) and Chi-square test all had 4 out of these 5 in their top 5. On the other hand, Random forest (permutation), Logistic regression (LIME), Deep networks (Permutation), Gaussian processes and NCA only had the PR interval of those listed in Table VI in their top 5 features.

TABLE VI
THE MOST COMMON FEATURES IN THE TOP 5 FOR 1^{st} DEGREE AV BLOCK FOR ALL 20 METHODS AND THEIR MODIFIED ROC AUC AND ROC AUC VALUES. FOR THE FREQUENCIES, WE PRESENT THE MEAN AND STANDARD DEVIATIONS AS INTEGERS. THESE VALUES ARE CALCULATED OVER 5 RUNS TO ACCOUNT FOR THE INHERENT RANDOMNESS OF THE METHODS.

Feature	Frequency in the top 5 features	Modified ROC AUC	ROC AUC
PR interval	20(0)	0.9384	0.9384
QRS duration	9(1)	0.7450	0.7450
T+ amplitude, lead I	8(1)	0.8247	0.1753
T morphology, lead I	6(1)	0.7289	0.2711
T+ amplitude, lead V6	6(1)	0.8175	0.1825

The ROC AUC values in Table VI indicate the direction of change of a feature with the pathology as described in Section II-D. Clearly, in this case, the PR interval increases with 1^{st} degree AV block, which is consistent with the cardiologists' diagnosis.

The QRS duration generally increases with 1^{st} degree AV block. The mean QRS duration for normal subjects is 92 ms which increases to 113 ms for 1^{st} degree AV block subjects. This is consistent with evidence of conduction slowing distal to the AV node in patients with known 1^{st} degree AV block.

The T+ amplitude in leads I and V6 decreases on average in patients with 1^{st} degree AV block according to these results. The physiological cause for these decreases is not clear.

Finally, the T morphology measure in lead I decreases with 1^{st} degree AV block, but this is an integer value representing different cases. Analysis of this feature shows that 99% of the values for the Normal category are +1, indicating a single upright T wave. However, for the 1^{st} degree AV block records, only 52% have a value of +1, with almost all the others having a value of either -1 or -2 in equal proportions. Thus, it seems that in approximately half the cases of 1^{st} degree AV block, the T wave is inverted or biphasic with negative leading component. A possible explanation for this is that for 1^{st} degree AV block subjects, the PR interval is longer resulting in a longer diastolic interval. If the action potential duration increases more in some regions than others for longer diastolic intervals (restitution), this could cause morphology changes in the T wave.

B. Right Bundle Branch Block

For RBBB, there are 7 important features and a further 3 features that correlate with at least one of these, as listed in Table I. Using the scoring algorithm described in Section II-E, we found the score for each method using the top 5 features of each ranking only. In Table VII, scores for each method comparing the top 5 features with both the important features and the important and correlating features are given. The best performing method is Logistic regression, while Random forest (SHAP and LIME), Random forest, Logistic regression (SHAP and LIME), Deep networks (SHAP), Chi-square test and Modified ROC AUC all have scores over 70. It is striking that only XGB (Permutation), Deep networks (LIME) and Gaussian Processes have an increased score when including the correlating features. The worst performing methods are

TABLE VII
RIGHT BUNDLE BRANCH BLOCK TOP 5 SCORES FOR THE DIFFERENT FEATURE IMPORTANCE RANKINGS USING THE IMPORTANT FEATURES ONLY OR THE IMPORTANT FEATURES TOGETHER WITH FEATURES THAT CORRELATE WITH THEM. THE RANKING OF THE NON-DISCRIMINATING FEATURES IS ALSO GIVEN. RESULTS FOR MODEL-DEPENDENT METHODS ARE GIVEN IN THE UPPER PART OF THE TABLE AND RESULTS FOR MODEL-INDEPENDENT METHODS ARE GIVEN IN THE LOWER PART. WE PRESENT THE MEAN AND STANDARD DEVIATIONS AS INTEGERS, CALCULATED OVER 5 RUNS, TO ACCOUNT FOR THE INHERENT RANDOMNESS OF THE METHODS.

Method	**Top 5 score important/ imp. + corr.**	**Ranking of the non-discriminating features**
Random Forest (permutation)	33(0)/33(0)	20(0)/61(0)
Random Forest (SHAP)	72(4)/72(4)	112(2)/115(0)
Random Forest (LIME)	82(14)/82(14)	110(0)/113(0)
Random Forest	79(7)/79(7)	113(1)/115(1)
XGB (permutation)	43(18)/47(19)	19(1)/56(1)
XGB (SHAP)	67(0)/67(0)	99(0)/106(0)
XGB (LIME)	38(5)/38(5)	110(0)/113(0)
Logistic Regression (permutation)	52(18)/52(18)	21(0)/69(6)
Logistic Regression (SHAP)	80(4)/80(4)	106(0)/78(0)
Logistic Regression (LIME)	82(6)/82(6)	109(0)/113(0)
Logistic Regression	**93(0)/93(0)**	76(2)/110(1)
Deep networks (permutation)	10(13)/10(13)	21(0)/1(0)
Deep networks (SHAP)	81(11)/81(11)	78(22)/34(13)
Deep networks (LIME)	28(18)/37(10)	109(0)/114(0)
Gaussian processes	60(0)/67(0)	4(0)/3(0)
Chi square test	87(0)/87(0)	116(0)/117(0)
MRMR	33(0)/33(0)	18(2)/2(0)
NCA	33(0)/33(0)	116(0)/117(0)
Relieff	60(0)/60(0)	106(0)/105(0)
Modified ROC AUC	73(0)/73(0)	116(0)/117(0)

Random forest (permutation), XGB (LIME), Deep networks (permutation and LIME), MRMR and NCA.

We then considered the ranking for each method of the non-discriminating features which are again shown in the final column of Table VII. We note that these feature rankings are rather low for Random Forest (SHAP and LIME), Random forest, XGB (SHAP and LIME), Logistic regression (SHAP and LIME), Deep networks (LIME), Chi-square test, NCA and ReliefF so all the SHAP and LIME methods do very well. However, these feature rankings are high for Random forest (Permutation), XGB (Permutation), Deep networks (Permutation), Gaussian processes and MRMR so all the Permutation methods perform poorly for these features. Deep networks (Permutation) even ranks one of the non-discriminant features as the most important one.

We again considered the top 5 features for each method, with the 5 most frequent shown in Table VIII. We note that 4 of these are important features. However, the S amplitude in lead V2 is not, but has a very high Modified ROC AUC value, ranking sixth in the ROC AUC ranking. Given that leads V1 and V2 are proximate on the body, it is not unexpected that the S amplitude in lead V2 is significant, mirroring the importance of the S amplitude in lead V1. The correlation coefficient between the two is reasonably high at 0.6707. The ROC AUC value suggests that the S amplitude in lead V1 increases with RBBB, resulting in a less pronounced S wave (as the S wave amplitudes are negative). All these features have a very high Modified ROC AUC value, which indicates good separation of the two distributions for these features, except for R' amplitude in lead V1.

Again, when averaging the rankings across all runs, three methods had all of their top 5 features in Table VIII, namely Logistic Regression, Logisitic Regression (LIME) and Deep networks (SHAP) while Random forest (LIME and SHAP), Random Forest, Logistic regression (SHAP), XGB (SHAP), Chi-square test and Modified ROC AUC all had 4 of their top 5 features in Table VIII. The worst performing methods were Logistic Regression (Permutation), Deep networks (permutation), and MRMR which had only one feature in Table VIII in their top 5.

The ROC AUC values in Table VIII show that QRS duration increases with RBBB, which is consistent with one of the diagnosis conditions that the width of the QRS complex should be >120 ms. The S amplitude in leads V1 and V2 increases with RBBB, resulting in shallower S waves since the S amplitude is negative, while the S amplitude in lead I decreases with RBBB, resulting in a deeper S wave. The R' amplitude in lead V1 increases with RBBB.

C. Left Bundle Branch Block

For LBBB, there are 14 important features and an additional 28 correlating features, as listed in Table II. The scoring algorithm described in Section II-E gives the scores as shown in Table IX, again using only the top 5 features. The scores for the important features only are generally quite low. However, when the correlating features are included, most methods show a significant improvement, which is not surprising as there are 28 additional correlating features, although much of the improvement in scores is due to the three T morphology features (see Table X).

Using only the important features, the best performing method is Gaussian processes, while Logistic regression (permutation) and Deep networks (permutation) both have a score of 0. When the correlating features are included, the Chi-square test has a perfect score of 100. Random Forest (SHAP and LIME), Random Forest, XGB (SHAP) and Relieff all have high scores above 90. In contrast, Logistic regression (permutation) still has a zero score while Deep networks (permutation) has an increased, but still poor, score of 5.

The rankings of the non-discriminating features were generally low, with Chi-square test and NCA performing particularly well. Gaussian Processes emerges as the only method that gives, with a ranking of 2, a particular high ranking to one of the non-discriminating features. In contrast, all the SHAP and LIME methods ranked this feature greater than 100, except for XGB (SHAP) which ranked it as 99, so these methods all performed well.

The frequency of features in the top 5 for all methods is shown in Table X. We note that three of these are correlating features, which may explain the big increase in scores when the correlating features are included. Again, all of these 5 features have a very high Modified ROC AUC value, indicating good separation of the two distributions for these features. We note that the three methods that did not have QRS duration in their top 5 features were Logistic regression (permutation) and Deep networks (permutation and LIME).

TABLE VIII
THE MOST COMMON FEATURES IN THE TOP 5 FOR RBBB FOR ALL 20 METHODS AND THEIR MODIFIED ROC AUC AND ROC AUC VALUES. FOR THE FREQUENCIES, WE PRESENT THE MEAN AND STANDARD DEVIATIONS AS INTEGERS. THESE VALUES ARE CALCULATED OVER 5 RUNS TO ACCOUNT FOR THE INHERENT RANDOMNESS OF THE METHODS.

Feature	**Frequency in the top 5 features**	**Type of feature**	**Modified ROC AUC**	**ROC AUC**
QRS duration	18(1)	Important	0.9933	0.9933
S amplitude, lead V1	11(1)	Important	0.9283	0.9283
R’ amplitude, lead V1	8(1)	Important	0.7860	0.7860
S amplitude, lead I	8(1)	Important	0.9234	0.0766
S amplitude, lead V2	7(1)	Unimportant	0.9199	0.9199

TABLE IX
LEFT BUNDLE BRANCH BLOCK TOP 5 SCORES FOR THE DIFFERENT FEATURE IMPORTANCE RANKINGS USING THE IMPORTANT FEATURES ONLY OR THE IMPORTANT FEATURES TOGETHER WITH FEATURES THAT CORRELATE WITH THEM. THE RANKING OF THE NON-DISCRIMINATING FEATURES IS ALSO GIVEN. RESULTS FOR MODEL-DEPENDENT METHODS ARE GIVEN IN THE UPPER PART OF THE TABLE AND RESULTS FOR MODEL-INDEPENDENT METHODS ARE GIVEN IN THE LOWER PART. WE PRESENT THE MEAN AND STANDARD DEVIATIONS AS INTEGERS, CALCULATED OVER 5 RUNS, TO ACCOUNT FOR THE INHERENT RANDOMNESS OF THE METHODS.

Method	**Top 5 score important/ imp. + corr.**	**Ranking of the non-discriminating features**
Random Forest (permutation)	33(0)/40(0)	56(0)/61(0)
Random Forest (SHAP)	33(0)/94(7)	77(12)/107(6)
Random Forest (LIME)	33(0)/92(7)	61(15)/112(0)
Random Forest	31(2)/97(3)	77(12)/98(11)
XGB (permutation)	33(0)/60(0)	56(0)/61(0)
XGB (SHAP)	33(0)/93(0)	59(0)/99(0)
XGB (LIME)	45(10)/61(19)	43(17)/112(0)
Logistic Regression (permutation)	0(0)/0(0)	56(0)/75(0)
Logistic Regression (SHAP)	33(0)/75(3)	40(1)/110(11)
Logistic Regression (LIME)	22(6)/66(10)	40(11)/111(1)
Logistic Regression	33(0)/60(0)	40(0)/53(0)
Deep networks (permutation)	0(0)/5(10)	56(0)/78(6)
Deep networks (SHAP)	15(10)/71(2)	62(11)/102(28)
Deep networks (LIME)	2(3)/66(18)	73(28)/110(1)
Gaussian processes	**60(0)**/67(0)	78(0)/2(0)
Chi square test	33(0)/**100(0)**	113(0)/117(0)
MRMR	33(0)/73(0)	116(0)/22(2)
NCA	40(0)/40(0)	100(0)/117(0)
Relieff	13(0)/93(0)	71(0)/101(0)
Modified ROC AUC	33(0)/80(0)	117(0), 116(0)

No method had all the top 5 features matching those in Table X but Random Forest (SHAP) had 4 out of its top 5 that matched with Table X. On the other hand, XGB (LIME) and NCA had only one of the features in Table X in their top 5, which was the QRS duration.

The ROC AUC values show that the QRS duration increases with LBBB, which is consistent with the condition that the width of the QRS complex should be >120 ms. The diagnosis of LBBB involves only changes in the QRS complex but the two T morphology features in Table X are not associated with the QRS complex. However, we have already noted they correlate strongly with the QRS duration.

The T morphology features for leads I and V6 decrease with LBBB. Analysis of these features shows that 99% of the values for the Normal class are +1 for both morphology features. For the LBBB records, 72% are -1 and 24% are -2 for the T morphology in lead I, and 69% are -1 and 24% are -2 for the T morphology in lead V6, both of which represent a significant shift from a single upright wave to either a single inverted wave or a biphasic wave with leading negative component.

The R amplitude in leads V3 and V4 are not important features for the diagnosis of LBBB, but this amplitude in leads V5 and V6 are important features. As leads V3 and V4 are very close to lead V5, it is not too surprising that these feature are common in the top 5 features for some methods. Interestingly, the R amplitude in leads V5 and V6 are not in the top 5 features for any method, so leads V3 and V4 seem to be more important than leads V5 and V6.

IV. COMPARISON WITH THE MULTICLASS CASE

We have considered feature importance ranking in the context of a binary classification of normal vs. a single pathology for three different pathologies, namely 1st degree AV block, RBBB and LBBB. This is the simplest possible case, but is not very realistic since cardiologists have to positively diagnose one (or more) conditions from a long list of possible conditions. It is also conceivable that a simple binary classification of normal vs. a specific pathology could be achieved with high accuracy using only a subset of the complete list of diagnostic conditions. Thus, as a next step, we considered feature importance ranking for a multiclass classification involving normal, 1st degree AV block, RBBB and LBBB records in [4]. The feature importance rankings were found for the one vs. all binary classifications as the aim is to positively diagnose one condition (since the data were single label) which implies a negative classification for the other classes.

The accuracies of the models were not reported in [4] but all four methods had an accuracy exceeding 95% for the multiclass classification. Also, the results for the model dependent methods are not directly comparable since the data were not normalised in [4] as they were in this study. In particular, the poor performance of Deep networks for the ranking of the PR interval for the 1st degree AV block case is almost certainly due to this lack of normalisation.

We now compare the feature rankings of the binary and multiclass cases.

A. First Degree AV Block

The ranking of the PR interval was very similar in the binary and multiclass cases. In the binary case, all methods ranked the PR interval as most important. In the multiclass case, the PR interval was not the top feature for Logistic regression (SHAP and LIME), Deep networks (SHAP and LIME) and Gaussian

TABLE X
THE MOST COMMON FEATURES IN THE TOP 5 FOR LBBB FOR ALL 20 METHODS AND THEIR MODIFIED ROC AUC AND ROC AUC VALUES. FOR THE FREQUENCIES, WE PRESENT THE MEAN AND STANDARD DEVIATIONS AS INTEGERS. THESE VALUES ARE CALCULATED OVER 5 RUNS TO ACCOUNT FOR THE INHERENT RANDOMNESS OF THE METHODS.

Feature	Frequency in the top 5 features	Type of feature	Modified ROC AUC	ROC AUC
QRS duration	17(0)	Important	0.9960	0.9960
T morphology, lead I	12(1)	Correlating	0.9689	0.0311
T morphology, lead V6	10(1)	Correlating	0.9510	0.0490
R amplitude, lead V4	5(1)	Correlating	0.9401	0.0599
R amplitude, lead V3	4(0)	Unimportant	0.9218	0.0782

processes. The poor results for Logistic regression and Deep networks are probably due to the fact that the data were not normalised.

The most common features in the top 5 had three features in common, namely the PR interval, QRS duration and T+ amplitude in lead I. The other features listed in Table VI are the T+ amplitude in lead V6 and T morphology in lead I whereas the other features for the multiclass case were the ST slope in leads I and V1 which are quite different features for the two cases.

B. RBBB

We first note that the correlating features for RBBB were different for the binary and multiclass cases, with 3 correlating features in the binary case (which are listed in Table I) and 5 correlating features for the multiclass case. The scores for the important and correlating features for the multiclass case are greater than the corresponding scores for the binary case for many methods, although a notable exception is Logistic regression (SHAP and LIME) which both had a score of zero in the multiclass case and were among the best scores in the binary case! We also note that in the multiclass case, the scores for the important and correlating features were 100 for four methods, namely Random forest, Random forest (permutation) and XGB (SHAP and LIME). The scores for MRMR and NCA were very low for the binary case, but improved significantly for the multiclass case, for which they had the second best score (important features only).

In this case, the most common features in the top 5 in the binary and multiclass cases have 4 features in common and so there is good agreement here.

C. LBBB

In this case, there are 28 correlating features in the binary case (which are given in Table II) but only 17 correlating features for the multiclass case. The scores for the important features only and for the important plus correlating features for the multiclass case were almost all less than the corresponding scores for the binary case.

The most common features in the top 5 only had no features in common in this case. The multiclass case includes the ST slope in three leads whereas the binary case includes the T morphology in two leads.

V. DISCUSSION

The results of the different feature ranking algorithms for the three pathologies that we have considered have some inconsistencies, although some general trends can be observed. For 1st degree AV block, all methods ranked the one important feature first. For RBBB, Logistic regression had the highest scores but scored quite poorly for LBBB. For LBBB, a score of 100 when including correlating features was obtained by Chi-square test, while Random forest (SHAP and LIME), Random forest, XGB (SHAP) and Relieff achieved high scores, almost reaching the perfect score. ReliefF performed poorly for LBBB (important features only) but had reasonable performance for RBBB.

If the scores for RBBB and LBBB are added together, then for the important features only, Logistic regression has the highest score, closely followed by Gaussian processes and Chi-square test. At the other end, Deep networks (permutation) has the lowest combined score. Adding the scores for RBBB and LBBB for the important and correlating features, then the top score is obtained by Chi-square test. It is closely followed by Random Forest, Random forest (SHAP and LIME) and Deep networks (SHAP). Meanwhile, the lowest combined score was obtained for Logistic regression (permutation) together with Deep networks (permutation).

When comparing the various methods combined with SHAP, LIME and permutation options, the permutation variations were consistently the worst. SHAP and LIME both produce comparably favorable outcomes, with SHAP exhibiting a slight overall advantage. However, the Random forest result for LBBB including correlating features was better than Random forest (SHAP and LIME) and Logistic regression results were significantly better than Logistic regression (SHAP and LIME) for RBBB. So the native feature importance rankings for Random forest and Logistic regression sometimes do well without the addition of other methods on top.

All of the SHAP (except in combination with Deep Nets) and LIME methods together with the Chi-square test and Random forest all ranked the non-discriminating features quite far down the rankings for RBBB and LBBB.

As mentioned in the introduction attribution methods applied to models operating on raw ECG data provide a complementary approach for knowledge discovery. However, a direct comparison of feature importance methods with the XAI attribution methods from [6] and [7] is not straightforward due to fundamental differences between both approaches. Feature importance methods typically focus on abstracted or constructed features, such as the ECG features in our research. While attribution methods can work with such derived features, as seen with SHAP in our work, their most compelling use case is with deep learning models that work on raw data. Attribution methods highlight the critical parts of the ECG that are relevant for predicting specific pathologies. However, even when results are aggregated across large groups of patients, as in [7], linking these relevant regions to the diagnostic concepts recognised by human cardiologists remains a challenge. While identifying aspects such as peak amplitudes from an attribution map might be straightforward, identifying intervals can be elusive. In contrast, feature importance methods provide a unified perspective by scoring individual ECG features. Given the strong differences in methodology, it can be difficult to draw parallels between the two. This complexity is compounded by the fact that the underlying models for each method are inherently different and may not agree on which features are considered important.

Therefore, it seems more appropriate to contrast the feature importance methods used for ECG features in this work with the concept-based explainability methods in [7], namely Testing with Concept Activation vectors (TCAV) [28]. As TCAV concepts are derived from ECG features or their aggregated combinations, they are more closely aligned with the conceptual domain of feature importance methods. Explainable AI methods were used in [7] to investigate concepts that are most relevant for the diagnosis of a number of conditions, including (complete) LBBB. In particular, TCAV was used to evaluate the importance of the concept "QRS complex exceeds 120ms" in the diagnosis of LBBB. Their results showed a statistically significant and strong correlation with this concept. This is consistent with our observations, where the QRS length feature was ranked highly. Despite the differences in the models and methods used - with [7] using convolutional neural networks on raw data - both studies consistently underlined the importance of QRS duration in LBBB prediction, in line with standard cardiologist guidelines. The synergy between different XAI/feature importance methods and model architectures deserves further investigation and represents an interesting avenue for future research.

VI. Conclusion

In this comparison of feature ranking algorithms with the expert knowledge of cardiologists for three different pathologies, we have shown that generally speaking, the SHAP and LIME methods all give good agreement with the important features used by cardiologists, together with the native Random forest and Logistic regression feature rankings. For the model independent methods, Chi-square test generally performed well. Some methods gave inconsistent results, including MRMR and NCA. The permutation methods generally performed quite poorly.

It is interesting that the top ranked features for many methods include some unimportant or correlating features rather than important features only. Notably, the T wave morphology features, which are conventionally not considered by clinicians, were consistently marked significant for the diagnosis of left bundle branch block.

The code for obtaining the feature importance rankings described in this work was made publicly available [18].

References

[1] "ECG Clinical Interpretation: A-Z by diagnosis," https://litfl.com/ecg-library/diagnosis/, accessed: 2022-05-09.

[2] B. Surawicz and T. K. Knilans, Eds., Chou's Electrocardiography in Clinical Practice, 6th ed. Saunders, 2008.

[3] C. Molnar, Interpretable Machine Learning: A Guide For Making Black Box Models Explainable, 2nd ed., 2022.

[4] P. J. Aston, T. Mehari, A. Bosnjakovic, P. M. Harris, A. Sundar, S. E. Williams, O. Dössel, A. Loewe, C. Nagel, and N. Strodthoff, "Multi-class ECG feature importance rankings: Cardiologists vs algorithms," Computing in Cardiology, vol. 49, 2022.

[5] H.-P. Schuster and H.-J. Trappe, EKG-Kurs für Isabel, 7th ed. Georg Thieme Verlag, 2017.

[6] Strodthoff, Nils and Wagner, Patrick and Schaeffter, Tobias and Samek, Wojciech, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," IEEE Journal of Biomedical and Health Informatics, vol. 25, pp. 1519–1528, 2021.

[7] Patrick Wagner and Temesgen Mehari and Wilhelm Haverkamp and Nils Strodthoff, "Explaining Deep Learning for ECG Analysis: Building Blocks for Auditing and Knowledge Discovery," Submitted, 2023.

[8] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," Scientific Data, vol. 7, no. 1, p. 154, 2020. [Online]. Available: https://doi.org/10.1038/s41597-020-0495-6

[9] P. Wagner, N. Strodthoff, R.-D. Bousseljot, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," PhysioNet, DOI 10.13026/x4td-x982., 2020.

[10] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215–e220, 2000.

[11] P. Macfarlane, B. Devine, and E. Clark, "The University of Glasgow (Uni-G) ECG analysis program," in Computers in Cardiology, 2005. IEEE, 2005, pp. 451–454.

[12] N. Strodthoff, T. Mehari, C. Nagel, P. Aston, A. Sundar, C. Graff, J. Kanters, W. Haverkamp, O. Dössel, A. Loewe, M. Bär, and T. Schaeffter, "PTB-XL+, a comprehensive electrocardiographic feature dataset," Scientific Data, vol. 10, p. 279, 2023.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[15] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," Journal of Machine Learning Research, vol. 20, no. 177, pp. 1–81, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-760.html

[16] R. Genuer and J.-M. Poggi, Random Forests with R. Springer International Publishing, 2020.

[17] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning. The MIT Press, 2005.

[18] T. Mehari, "ECG Feature Importance Rankings: Cardiologists vs. Algorithms: Codebase," https://github.com/tmehari/feature_importance, 2023.

[19] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Feature selection for high-dimensional data," Progress in Artificial Intelligence, vol. 5, no. 2, pp. 65–75, Feb. 2016. [Online]. Available: https://doi.org/10.1007/s13748-015-0080-y

[20] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," Computers in Biology and Medicine, vol. 112, p. 103375, Sep. 2019. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2019.103375

[21] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," Information Fusion, vol. 52, pp. 1–12, Dec. 2019. [Online]. Available: https://doi.org/10.1016/j.inffus.2018.11.008

[22] C. D. Manning, P. Raghavan, and H. Schütze, An Introduction to Information Retrieval. CUP, 2008.

[23] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” J. Bioinform. Comput. Biol., vol. 3, pp. 185–205, 2005.

[24] Z. Zhenyu, R. Anand, and M. Wang, “Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform,” in International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2019, pp. 442–452.

[25] Y. Wei, K. Wang, and W. Zuo, “Neighborhood component feature selection for high-dimensional data,” Journal of Computers, vol. 7, no. 1, pp. 161–168, 2012.

[26] I. Kononenko, “Estimating attributes: Analysis and extensions of RELIEF,” in Machine Learning: ECML-94. Springer Berlin Heidelberg, 1994, pp. 171–182. [Online]. Available: https://doi.org/10.1007/3-540-57868-4_57

[27] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, pp. 861–874, 2006.

[28] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas et al., “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav),” in International conference on machine learning. PMLR, 2018, pp. 2668–2677.

APPENDIX A
ECG FEATURES

The 117 features from the Glasgow 12-lead ECG analysis algorithm [11] that we identified as ones that cardiologists would typically consider when making a diagnosis mainly consist of features derived for all 12 leads, which are as follows:

- Peak-to-peak amplitude
- Q amplitude
- R amplitude
- S amplitude
- R’ amplitude (amplitude of a second R peak)
- T+ amplitude (maximum height of the T wave)
- P morphology
- T morphology
- ST slope

The morphology parameters are integers representing four cases, namely:

- A biphasic wave with leading positive component (+2)
- A single upright wave (+1)
- A single inverted wave (-1)
- A biphasic wave with leading negative component (-2)

In addition, a number of measurements derived from all 12 leads were used as follows:

- QRS frontal axis
- Average RR interval
- Heart rate variability
- Overall ST duration
- Overall PR interval
- QTc (Framingham)
- Overall P duration
- Overall QRS duration
- Overall T duration