applied sciences

MDPI

Article

Perturbation-Based Explainable AI for ECG Sensor Data

Ján Paralič [1,*], Michal Kolárik [1], Zuzana Paraličová [2], Oliver Lohaj [1] and Adam Jozefík [1]

[1] Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Letna 9, 040 01 Košice, Slovakia
[2] Faculty of Medicine, University of Pavol Jozef Šafárik, Trieda SNP 1, 040 11 Košice, Slovakia
* Correspondence: jan.paralic@tuke.sk

Abstract: Deep neural network models have produced significant results in solving various challenging tasks, including medical diagnostics. To increase the credibility of these black-box models in the eyes of doctors, it is necessary to focus on their explainability. Several papers have been published combining deep learning methods with selected types of explainability methods, usually aimed at analyzing medical image data, including ECG images. The ECG is specific because its image representation is only a secondary visualization of stream data from sensors. However, explainability methods for stream data are rarely investigated. Therefore, in this article we focus on the explainability of black-box models for stream data from 12-lead ECG. We designed and implemented a perturbation explainability method and verified it in a user study on a group of medical students with experience in ECG tagging in their final years of study. The results demonstrate the suitability of the proposed method, as well as the importance of including multiple data sources in the diagnostic process.

Keywords: deep learning; explainable AI; ECG signals; perturbation method

check for updates

Citation: Paralič, J.; Kolárik, M.; Paraličová, Z.; Lohaj, O.; Jozefík, A. Perturbation-Based Explainable AI for ECG Sensor Data. *Appl. Sci.* **2023**, *13*, 1805. https://doi.org/10.3390/app13031805

Academic Editors: Yang-Lang Chang, Mohammad Alkhaleefah and Tan-Hsu Tan

Received: 13 December 2022
Revised: 17 January 2023
Accepted: 27 January 2023
Published: 31 January 2023

1. Introduction

Electrocardiography (ECG) is one of the most essential methods of examination in internal medicine. It is a non-invasive diagnostic method that records the electrical activity of the heart from the surface of the body. Electrodes placed on the skin measure the electrical potential differences that occur during the spread of the action potential through the myocardium. The ECG device then graphically records these changes as an ECG curve. According to the nature of the curve, it is possible to evaluate heart rate, heart rhythm, and many pathological conditions, including myocardial infarction. The correct evaluation of the ECG curve is crucial for determining the correct diagnosis, which often requires urgent treatment. Artificial intelligence methods could contribute to eliminating human error and to obtaining faster and more accurate diagnoses. Indeed, several models of artificial intelligence, especially deep neural networks, have produced significant results in solving various challenging tasks, including medical diagnostics. However, these complex models have the character of a "black box", meaning that, as a rule, they cannot provide a comprehensible explanation of the proposed decision for the doctor to assess. Therefore, in order to increase the credibility of these black-box models in the eyes of doctors, it is necessary to focus on their explainability.

This article addresses the question of the explainability of artificial intelligence models of the deep neural network type for the diagnosis task on stream data from 12-lead ECG. For this purpose, in Section 2, we analyze relevant published papers focused on diagnostic models of ECG data which address the issue of explainability. As this analysis showed, most authors have trained deep neural network models with different architectures on an image representation of ECG data. Therefore, we focused on modeling ECG data in their initial form, in which they have the character of stream sensor data. The primary aim of the research presented in this article was to design a specific perturbation-based method to explain deep learning models trained on ECG signals, then verify the results

by comparison to experienced medical students. In Section 3, we describe the dataset we used along with its preprocessing. In Section 4, we present the resulting model based on a convolutional neural network, which was trained on the available data. The main contribution is presented in Section 5, where the original explainability method is proposed and verified in a user study. Finally, the most significant contributions and open research problems are summarized in Section 6.

2. Methods and Models for Processing and Explaining ECG Data

ECG data belong to the category of time series data. Rojat et al. [1] developed a survey concerning explainable artificial intelligence on this data type. They talked about time series as an area of research that has not received as much attention as computer vision and language processing. The same view was presented by Simic et al. [2] in their review paper, in which they discussed XAI methods for time series data. However, the lack of explainability is a significant disadvantage in the medical field, which motivated our own research.

In analyzing trends, we based our research on the current application of explainability methods to ECG data, as described in several papers. Vijayarangan et al. [3] used deep learning methods for heart arrhythmia classification in their work. Their paper describes an approach using a traditional convolutional neural network and the GradCAM saliency method to explain predictions. The second approach used an LSTM network in which saliency was derived from the input erasure mask during model learning. They compared their visualizations with baseline values, and their results were consistent with the medical literature on cardiac arrhythmia classification. Strodthoff et al. [4] focused on detecting and interpreting myocardial infarction using a fully convolutional neural network. Their classification reached 93.3% sensitivity and 89.7% specificity in ten-fold cross-validation. For interpretation, they used the Gradient*Input method, while in the output they emphasized only the blue and red areas of the ECG curves, i.e., those that did or did not contribute to the correct assessment of myocardial infarction. Raza et al. [5] created a complex model for classifying different kinds of cardiac arrhythmia from ECG data. They used a convolutional neural network to create an auto-encoder and a convolutional network for direct data classification. They achieved a precision rate of 94.5% and 98.9% in cardiac arrhythmia classification using noisy and cleaned data with five-fold cross-validation. For clarity, they used the GradCAM method in combination with the auto-encoder output to highlight the areas of the ECG waveform responsible for the model's decision.

An overview of XAI methods for multiple types of data, including sensory data and ECG data, was provided by Jeyakumar et al. [6] in their study. In an empirical study, they compared popular XAI methods with each other concerning user provenance on different types of data. The authors wanted to show the methods' universality and to determine which method was suitable for which type of data. They used the ECG dataset for sensory data, and compared the Grad-CAM++, Saliency Maps, SHAP, and Explanation by Example methods. In the case of ECG data, Explanation by Example was the most suitable method thanks to its ease of understanding and lower user uncertainty, as Saliency methods may highlight flat sections instead of actual heartbeat spikes. On the other hand, methods highlighting the most important regions of the model are able to alert the physician to a potential problem rather than displaying a similar image from the training set.

All of the mentioned publications achieved interpretation by applying explainability methods after neural network training (i.e., post hoc methods). However, a number of these methods were designed for other data types, such as table or image data, then made model-agnostic for any model architecture. The LIME, SHAP, or Explanation by Example methods can be used for various data types. Rojat et al. [1] divided post hoc methods to explain time series using convolutional neural networks into the following two categories.

The first category, Backpropagation methods, provide an explanation by moving forward and backward in a neural network. They were originally designed for computer vision. These methods are dominant in ECG analysis, and were used in the aforementioned

papers. The most commonly used applications are class activation mapping methods (CAM and Grad-CAM) and Saliency Maps.

The second category, Perturbation methods, use a modification of the model input, which can be an image, words in the text, an attribute in a table, or another type of data, while capturing changes in the output. The more significant the change in the output, the more partially relevant the input is to the model. The most significant changes occur if the target class of the sample on input changes. The biggest challenge with perturbation-based methods is the number of possible combinations. Suppose we want to go through all the elements on the input and obtain the corresponding output for them; one approach could be to select a plausible group of elements on the input and test perturbations within this group to obtain an optimal solution approximation, for example, choosing an appropriate window width for time-based sequences. The advantages of perturbation models over other models are that they are simple, allow for iterative querying and testing of hypotheses on the fly, and independent of the model architecture, unlike Saliency methods, which require internal model information. On the other hand, perturbation methods are computationally expensive, especially when using high-dimensional data, and may not be able to detect interactions between features used by the model. The output of perturbation methods consists of the ability to visualize and highlight the important parts of the input data, for example, a portion of an image or text. Ivanovs et al. [7] produced a survey paper devoted to the application of perturbation-based methods to different types of data, such as images, videos, and text.

In their work, Schlegel et al. [8] compared the quality of selected XAI methods by changing the verification method. For verification, they used perturbation of the input data over time. First, they performed a perturbation on the initialization value for t = 0 and the inverse (previous) value in time. For the second verification approach, they used perturbation of a time sequence instead of a point in time. Then, the perturbation method was taken as the ground truth when testing methods such as Saliency, LRP, DeepLIFT, LIME, and SHAP. In the obtained results, the SHAP method outperformed all other methods.

Perturbation methods for ECG data detect the curve's influence on the neural network's output by erasing, masking, or changing part of the ECG. Subsequently, they pass on the new input to the neural network, which divides it into classes. The results of the neural network's prediction on the old and new curves are compared. The higher the difference in class prediction, the more important the curve sequence that has been altered. Perturbation methods are applicable as long as the distance between the outputs of the neural network can be calculated. A suitable distance metric is, for example, the probability that a given input represents a given class. While these methods are not as popular, they offer a more straightforward implementation and understanding of the interpretation of the results.

3. Dataset and Data Preparation

The most suitable dataset comes from the public competition CinC 2021 (Computing in Cardiology Challenge 202113). This a dataset contains 10,344 ECG-labeled records originating from the southeastern United States. Each recording is 5 to 10 seconds long with a sampling frequency of 500 Hz. Recordings are in 12-lead ECG format. The organizers encouraged participants to use open-source licenses for their algorithms. This kind of license means that there are freely available models for predicting this data. The data are in WFDB format, which contains saved ECG records as MATLAB V4 binary files. There is a library (package), WFDB15, for the Python language to save, write, and process the data.

In the competition results, Reyna et al. [9] state that more than 200 teams presented neural networks focused on ECG data. Features for the data processing and deep architecture of the neural networks were used from the publicly available Singstad and Tronstad [10] competitive solution. Singstad and Tronstad used data in the same format as the freely available dataset used in this paper. In their work, they examined eight black-box type models. Based on the area under the precision–recall curve (AUPRC) metric, the best-rated

model was an auto-encoder neural network combined with a fully convolutional neural network with added rule-based algorithms.

Each patient record is contained in two files. One contains ECG signals, while the second is a header containing metadata such as gender, age, and diagnosis. All diagnoses are encoded by SNOMED-CT code (https://www.nlm.nih.gov/healthit/snomedct/index.html, accessed on 7 May 2022). Therefore, functions for processing data were necessary for working with the data files. Data preprocessing consisted of loading metadata and ECG recordings into vectors. Subsequently, it was necessary to transform all nominal attributes into numerical ones, including for sex and age and to handle missing values and coding of diagnoses.

After processing the data using the functions from the competition example, it was possible to obtain the statistical values for diagnoses. Signals were displayed using loading and rendering packages for ECG data. The age of the patients was normally distributed, with a peak between 60–80 years. The dataset contained various combinations of 24 diagnoses. In total, it contained 830 different combinations of these diagnoses. In Figure 1, the number of cases for individual diagnoses is displayed. As can be seen, the most frequently occurring diagnoses were: deviated T-bridge, sinus rhythm, and bradycardia.

Dx	SNOMED CT Code	Abbreviation	Total cases
1st degree av block	270492004	IAVB	769
atrial fibrillation	164889003	AF	570
atrial flutter	164890007	AFL	186
bradycardia	426627000	Brady	6
complete right bundle branch block	713427006	CRBBB	28
incomplete right bundle branch block	713426002	IRBBB	407
left anterior fascicular block	445118002	LAnFB	180
low qrs voltages	251146004	LQRSV	374
nonspecific intraventricular conduction disorder	698252002	NSIVCB	203
pacing rhythm	10370003	PR	0
premature atrial contraction	284470004	PAC	639
premature ventricular contractions	427172004	PVC	0
prolonged pr interval	164947007	LPR	0
prolonged qt interval	111975006	LQT	1391
qwave abnormal	164917005	QAb	464
right axis deviation	47665007	RAD	83
right bundle branch block	59118001	RBBB	542
sinus arrhythmia	427393009	SA	455
sinus bradycardia	426177001	SB	1677
sinus rhythm	426783006	SNR	1752
sinus tachycardia	427084000	STach	1261
supraventricular premature beats	63593006	SVPB	1
t wave abnormal	164934002	TAb	2306
t wave inversion	59931005	TInv	812
ventricular premature beats	17338001	VPB	357

Figure 1. Particular diagnoses and the respective number of cases for individual diagnoses in the dataset.

For training, it was necessary to adjust the examples to a uniform size. The number of signals within individual records ranged from 5000 to 10,000. More than 90% of the examples had a length of 5000. Therefore, the uniform size was set to this value. Records that were longer were abbreviated. The examples prepared this way were divided into

training and validation sets by stratification, ensuring that the same percentage of examples from each class was preserved.

4. Model Training and Evaluation

Competing solutions used different neural network architectures to train models, for example, residual neural networks, convolutional neural networks, and auto-encoder neural networks. In addition, a number of networks used age and gender attributes along with ECG for training.

We used a simple convolutional neural network for training because our primary focus was on explainability. It was necessary to use a simple neural network for learning because a complex neural network increases the computational time of the explainability method. The reason for this is that post hoc explainability methods measure the distance between predictions. Based on the AUPRC metric, Singstad and Tronstad [10] found that a convolutional neural network achieves the best results among simple architectures.

CNN networks are primarily designed for image classification problems, where the model can learn features from a two-dimensional input. This process can be problematic for one-dimensional data sequences such as sensory data. A 1D CNN model can learn an internal representation of time-series data without manually engineering input features, as in other ML models. The dataset used here has twelve features, with a window size of 5000 time steps (signals). For implementation, we used the Keras deep learning library, with which the input for a 1D CNN requires a three-dimensional input (samples, time steps, and features). There were three convolutional layers between the input and output layers; the convolutional layers used ReLU activation, and batch normalization was used for data batches. The number of neurons is summarized in Figure 2.

The respective number of output filters for the convolution layers was 8, 5, and 3. The output layer used a sigmoid activation function that assigned a probability value based on the likelihood of the example being in the respective class to each of the 24 attributes. The error function was set to BinaryCrossentropy, and the optimizer was set to Adam. The model was trained on the data in batches of 30 examples. After several iterations, the number of epochs was set to 10.

Evaluation

During training, the algorithm was set to monitor the accuracy, recall, and precision. These metrics were tracked over ten epochs. For evaluation of the model, a confusion matrix was used to display the resulting predictions on the validation set. One ECG record could be marked with several arrhythmias. A neural network prediction contains a probability for each target attribute value. The key is to learn a threshold that accurately matches the target attribute's value.

The threshold was calculated using the same method that Singstad and Tronstad [10] presented in their work. The classifier ran on the training data and received a score between 0 and 1 for each class. The Nelder–Mead downhill simplex method [11,12] was used to find the optimal threshold for all 24 classes. A prediction value for a target attribute greater than threshold = 0.4899078 was considered true, while a value less than or equal to the threshold was considered false. The confusion matrix is shown in Figure 3.

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 5000, 12)]        0

 conv1d (Conv1D)             (None, 5000, 128)         12416

 batch_normalization (BatchN  (None, 5000, 128)        512
 ormalization)

 activation (Activation)     (None, 5000, 128)         0

 conv1d_1 (Conv1D)           (None, 5000, 256)         164096

 batch_normalization_1 (Batc  (None, 5000, 256)        1024
 hNormalization)

 activation_1 (Activation)   (None, 5000, 256)         0

 conv1d_2 (Conv1D)           (None, 5000, 128)         98432

 batch_normalization_2 (Batc  (None, 5000, 128)        512
 hNormalization)

 activation_2 (Activation)   (None, 5000, 128)         0

 global_average_pooling1d (G  (None, 128)              0
 lobalAveragePooling1D)

 dense (Dense)               (None, 24)                3096

=================================================================
Total params: 280,088
Trainable params: 279,064
Non-trainable params: 1,024
_________________________________________________________________
None
```

Figure 2. Summary of the neural network model used for classification of 12-lead ECG signals into 24 categories.

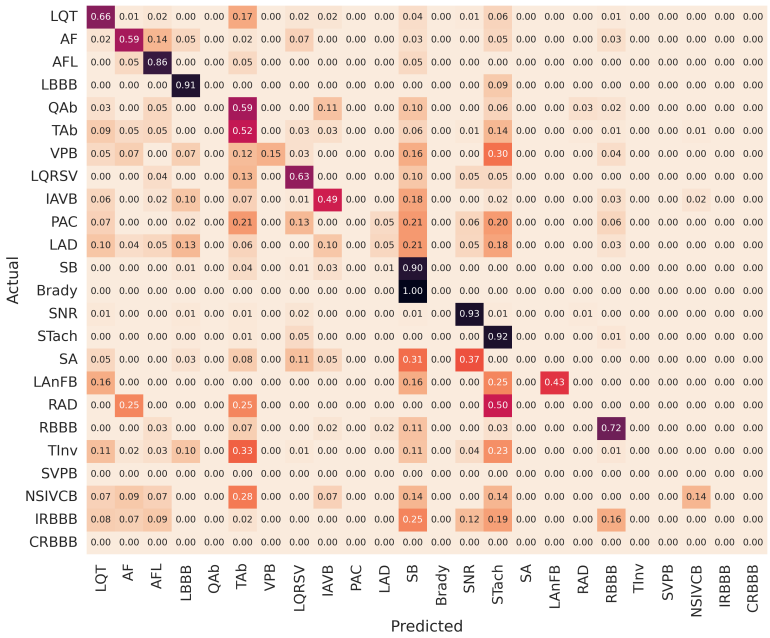


Figure 3. Confusion matrix on the validation set. Abbreviations on left in rows and bottom in columns correspond to particular diagnoses (24 categories). Numbers present the distribution of cases of the actual class classified into particular predicted classes. The sum in each row is 1.

In addition to the confusion matrix, the target attribute values were evaluated using the accuracy, recall, and F1-score metrics. In Figure 3, the resulting metrics on the validation set are displayed for individual target classes. Summarizing metrics are labeled according to the type of averaging used as "micro" and "macro". Macro-averaging assigns equal weights to each class, while micro-averaging assigns equal weights to each sample. When the number of examples in particular classes is approximately the same, both are acceptable. However, for imbalanced data, as in our case, which approach is preferable depends on the application. In our case, all diagnoses are important, regardless of whether they are frequent or rare; therefore, macro-averaging is preferable. The recall metric is very important if true positive cases need to be correctly predicted. According to the recall metric, the model best classified the values of the target attributes SB, STach, and LBBB. Our primary goal, however, was not to produce the best model; rather, it was to evaluate our new perturbation-based explanation method on a model of acceptable quality. Alternatively, the more sophisticated models presented in [10] achieved a macro-average F1-score of 0.3–0.4. Our simple convolutional neural network achieved an F1 score of 0.27. Compared to current models, although our model achieved slightly worse results, it saved significant computing time with the implemented explainability method, which was our primary goal.

5. Model Explanations

In [1], the authors of the survey developed a summary of XAI methods applied to time series. Out of 35 explainability methods, only three were expert-oriented. Combining domain knowledge with explainable artificial intelligence involves challenges such as explainability quantification, performance trade-offs, and information loss. The authors in [13] found that, in their experiments, the information loss was very negligible. Furthermore, the explainability score was better when domain knowledge was used than when no knowledge was used. We consulted the process of design, implementation, and verification of the explainability method with a fifth-year student of the Faculty of Medicine. This student had two years of experience in a hospital and participated in a study where he was tasked with tagging ECG recordings.

5.1. Designed Method

Available libraries and tools do not support ECG data and sensor data. Open-source code on this topic is not available, and the existing use cases devote very little space to explainability. Therefore, we decided to implement our own explainability algorithm. Current solutions of interpretability methods on ECG data highlight important parts of the image visualizing the ECG recording. However, such a solution poses a number of unanswered questions, e.g., the optimal length of the segment for which importance is assigned, the optimal way to assign importance to a given segment, and how the explainability method is affected by the target group of users (developer/expert).

We have attempted to answer these questions in the algorithm we propose for the explainability method based on the perturbation principle. The principle of perturbation methods is as follows:

1. Calculate the prediction of the original example for a certain diagnosis.
2. Remove, mask, or alter part of the sensory data. In our case, values in the selected interval to be perturbed are set to zero.
3. Pass the new input to the black-box model and measure the difference between the resulting examples.
4. In the case that the accuracy of the model is significantly reduced after perturbation, the changed part of the sensory data are important; assign importance to the changed part accordingly.

The pseudocode of the implemented perturbation method for an individual ECG lead is provided in Algorithm 1.

Algorithm 1 XAI for individual lead

Procedure: XAI_FOR_INVIDUAL_LEAD (*Ecg_Instance*, *NN_Model*, *Size_Of_Perturbed_Interval*, *Lead*, *Syndrom_Of_Interest*)
Input: Instance of ECG data *Ecg_Instance*, **Neural network model** *NN_Model*, **Size of perturbed interval** *Size_Of_Perturbed_Interval*, **Lead** *Lead*, **Syndrom of Interest** *Syndrom_of_Interest*
Output: Importance of intervals *Importance_Of_Intervals*

Prediction ← NN_Model.predict(*Ecg_Instance*);
initialize *Number_Of_Intervals* ← *length*(*Ecg_Instance*[*Lead*])/ *Size_Of_Perturbed_Interval*
initialize *Importance_Of_Intervals*[]
for $i = 0, \ldots,$ *Number_Of_Intervals* **do**
    New_Ecg ← *Perturb*(*Ecg_Instance*[*Lead*][$(i * Size\_Of\_perturbed\_Interval) : ((i+1) * Size\_Of\_perturbed\_Interval)$]
    Prediction_for_Perturbed_instance ← *NN_Model.predict*(*New_Ecg*)
    Importance_Of_intervals[*i*] ← *Prediction*[*Syndrom_Of_Interest*] - *Prediction_for_Perturbed_instance*[*Syndrom_Of_Interest*]
end for
return *Importance_Of_Intervals*;

An important aspect of explainability is its good visualization. After obtaining the values of the intervals from the function XAI_FOR_INVIDUAL_LEAD, it is necessary to assign colors from the heatmap to these values. A heat map is a map that distinguishes the most important places on the ECG recording by color. The heatmap spectrum contains colors: dark red, red, orange, and white, where dark red represents the most important area and white is the unimportant area. The heat map is plotted on the ECG curve. The ECG curve must meet certain characteristics in order to be easily readable by an expert. The basis for ease of reading is the grid, which is drawn below the ECG waves. The grid must be sufficiently specific and its parts must be clearly defined. We defined the grid as follows: one interval on the x-axis represents 0.2 s and one interval on the y-axis describes 50 microvolts. We have defined the degree of detail to ensure that it is as similar as possible to the examples according to [14], from which respondents learn to identify syndromes. An example of a graph that shows explainability is provided in Figure 4.

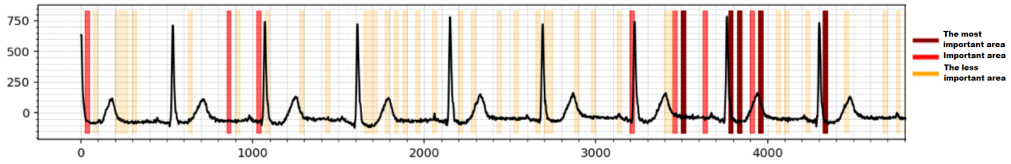


Figure 4. An example of an ECG graph that presents explanations in form of heatmap coloring.

The XAI_FOR_INVIDUAL_LEAD function has a parameter Size_Of_Perturbed_Interval. Setting this parameter specifies the time window to be perturbed. After perturbation, the function creates a prediction based on a new record. Decreasing the size of this window increases the computation time of the function, while increasing the size of this window may cause information loss. Based on several experiments, Size_Of_Perturbed_Interval = 0.05 sec turned out to be the best value. This value achieved sufficient specificity and a short calculation time for the main function.

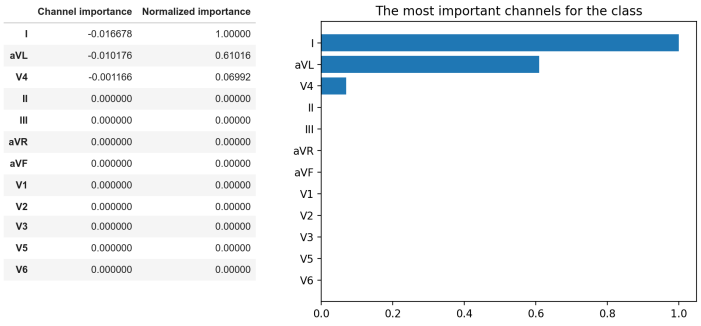
As we were working with a 12-lead ECG recording, the question arose of which lead or leads to choose in order to provide the best visualization and explanation to the doctor. For this task, we implemented the approach proposed by the authors in [15]. The function MOST_IMPORTANT_LEAD determines the most important ECG lead for the prediction of the given syndrome. The pseudocode is presented in Algorithm 2.

Algorithm 2 Most important lead

Procedure: MOST_IMPORTANT_LEAD (*Ecg_Instance*, *NN_Model*, *Syndrom_Of_Interest*)
Input: Instance of ECG data *Ecg_Instance*, **Neural network model** *NN_Model*, **Syndrom of Interest** *Syndrom_of_Interest*
Output: Lead importance *Lead_Importance*

Prediction ← NN_Model.predict(*Ecg_Instance*);
initialize *Lead_Importance*[]
for $i = 0, \ldots, 12$ **do**
    New_Ecg ← *Perturb*(*Ecg_Instance*[*i*])
    Prediction_for_Perturbed_instance ← *NN_Model.predict*(*New_Ecg*)
    Lead_Importance[*i*] ← *Prediction*[*Syndrom_Of_Interest*] - *Prediction_for_Perturbed_instance*[*Syndrom_Of_Interest*]
end for
return *Global_predictions*;

The output of the function MOST_IMPORTANT_LEAD is a list of values. These values describe the degree of reduction in the prediction of the specified syndrome when changing the individual leads. The greater the reduction due to the resulting lead perturbation, the more important it is to the prediction. The magnitudes of changes vary for individual syndromes. Therefore, we added normalized importance and a horizontal bar graph for clearer quantification of the importance. An example of the output of the MOST_IMPORTANT_LEAD function is shown in Figure 5, along with representation of the most important leads for the prediction of long QT syndrome.



	Channel importance	Normalized importance
I	-0.016678	1.00000
aVL	-0.010176	0.61016
V4	-0.001166	0.06992
II	0.000000	0.00000
III	0.000000	0.00000
aVR	0.000000	0.00000
aVF	0.000000	0.00000
V1	0.000000	0.00000
V2	0.000000	0.00000
V3	0.000000	0.00000
V5	0.000000	0.00000
V6	0.000000	0.00000

Figure 5. An example of the output of the MOST_IMPORTANT_LEAD function. Representation of the most important leads for the prediction of long QT syndrome is visualized in the form of a table and a bar plot.

5.2. Evaluation and Discussion

Based on communication with experts, we selected a subset of diagnoses from the data. This subset had to meet several conditions. The selected diagnosis had to have a sufficient number of cases (more than 400) in the data; at the same time, the quantitative metrics of selected diagnoses had to meet sufficient values (F1-score greater than 0.4, accuracy greater than 0.6). Experts had to be able to identify and know the given diagnoses. Thus, the selected diagnoses had to have been included in the study materials of the selected students. Diagnoses described in the recommended study materials for medical students according to [14] were the following:

- Sinusoidal rhythm (SNR)
- Sinus bradycardia (SB)
- Sinus tachycardia (STach)
- Abnormal T wave (TAb)
- Long QT syndrome (LQT)
- Atrial fibrillation (AF)
- Blockade of the right Tawar branch (RBBB)
- Lowered amplitude of the QRS complex (LQRSV)
- First-degree AV block (IAVB).

Verification was based on the assumption that if the method correctly provides explanations for a subset of the target attribute values, it will correctly provide it for the rest. The correctness of the explainability methods was verified using a short questionnaire and an interview with the respondents. The questionnaire contained a series of ECG recordings with the XAI_FOR_INVIDUAL_LEAD interpretability method applied. The method of explainability of the most important lead is based on the assumption that certain ECG leads are more important for predicting certain diagnoses. The reason for this is that diagnoses are based on certain parts of the ECG wave. Each ECG lead looks at the heart from a different angle and sees the part of the heart that is closest to the lead. The verification process was as follows:

1. Calculate the most important lead using the MOST_IMPORTANT_LEAD function on all cases of selected diagnoses in the test data. Only one most important lead is

selected for one case. The number of cases in which individual leads were the most important for each diagnosis is counted.

2. The relevance of identified most important leads with individual diagnoses has been confirmed in most of the cases. Connections of the most important leads with diagnoses were acknowledged based on communication with the expert:
    - Sinus bradycardia (SB)—U wave in leads (aVR, V2, V4–6) may be present in bradycardia.
    - Sinus tachycardia (STach)/Long QT syndrome (LQT)—P waves positive in leads I and II (these are the basic sign of sinus rhythm, which is important for the diagnosis of tachycardia).
    - Abnormal T wave (TAb)—Low QRS voltage negative inverted T waves are often seen in lead III.
    - Atrial fibrillation (AF)—ST depression and negative T waves in the lateral leads (I, aVL, V5, V6), Ashman's phenomenon (V5 lead).
    - Blockade of the right Tawar branch (RBBB)—RBBB is best "looked" at through leads V1–V5.
    - First-degree AV block (IAVB)—leads I and II are used for diagnosis, as they can detect a positive Physiological P wave.
3. The relevance of the identified most important leads was refuted in the following cases based on communication with the expert:
    - Sinus Rhythm (SNR)/Sinus Bradycardia (SB)/Sinus Tachycardia (STach)/Long QT Syndrome (LQT)—generally, Lead II is considered the best lead for finding atrial activity.
    - Lowered amplitude of the QRS complex (LQRSV)—AVL lead is not related, as it uses the Sokolow index for diagnosis, which uses leads (V1/V2/V5/V6).

Based on the confirmed connections of the leads with individual diagnoses, the correctness of the explainability method was confirmed. The MOST_IMPORTANT_LEAD function considered the same leads to be important as the experts in the given area. The exceptions were the diagnoses LQT, LQRSV, SNR, for which the most important lead was not suitable for diagnosis; however, in these cases the next lead identified by the MOST_IMPORTANT_LEAD function already fulfilled the suitability criteria.

The function XAI_FOR_INVIDUAL_LEAD marks the most important parts (Figure 4) of a specific ECG lead for the prediction of a specific diagnosis. The diagnoses in the data are of different natures. In general, the diagnoses can be divided into:

- Diagnoses describing the length (time period) of waves and oscillations
- Diagnoses describing the height (amplitude) of waves and oscillations
- Diagnoses describing the shape of waves and oscillations
- Diagnoses describing basic heart rhythms.

Sinus bradycardia (SB)/Sinus tachycardia (STach) are attributes that describe low/high heart rates. Atrial fibrillation is a diagnosis that refers to the irregular action of the heart. For such diagnoses, the expert looks at the entire stream. The XAI_FOR_INVIDUAL_LEAD function did not make sense for SB/STach/SR diagnoses, because the SB/STach/SR diagnoses did not describe specific parts of the ECG recordings.

When initially communicating with experts, asking about the most important areas caused confusion. The task of marking important areas was too vague. For verification, we therefore asked medical students to what extent the red (important) areas indicate specific waves and oscillations. For example, for a diagnosis that indicated an abnormal T wave, we asked: "To what extent do the red areas indicate a T wave?". Figure 6 shows examples of correct and incorrect R wave labeling. The left side of the image shows the R wave. The middle part of the figure is an example where the algorithm incorrectly labeled the R wave. The right side of the figure shows the correct R wave label.

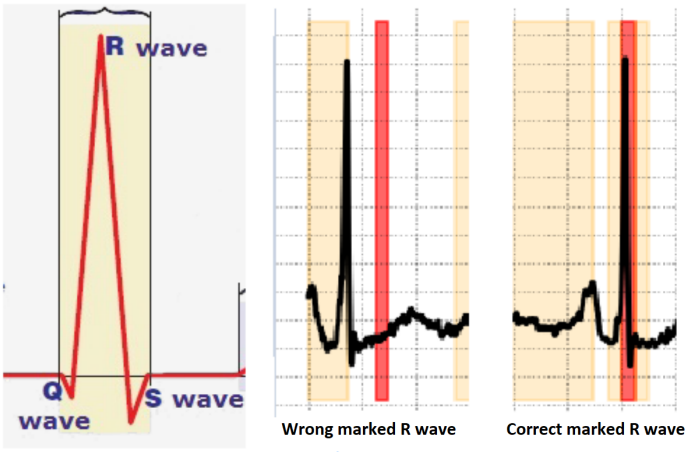


Figure 6. An example of correct and incorrect R wave labeling: the left side of the image shows the R wave; the middle part of the figure shows an example where the algorithm incorrectly labeled the R wave; the right side of the figure shows the correct R wave label.

Verification was carried out using a questionnaire. The questionnaire contained five graphs along with explanations provided by our algorithm. The respondents' task was to determine, on a scale of 1 to 5 (1 = completely wrong, 5 = completely right), to what extent the red areas indicated a wave associated with a particular diagnosis. For example, the image in Figure 7 zooms in on one such graph with explanation, and shows the question asked and the summary results of the respondents. The results of individual students were very similar. Responses for all graphs ranged from grade 2 to grade 4 (average grade 3). It took the respondents ten minutes on average to rate the five graphs. The XAI_FOR_INVIDUAL_LEAD function successfully marked those parts of the areas important for the given diagnoses.

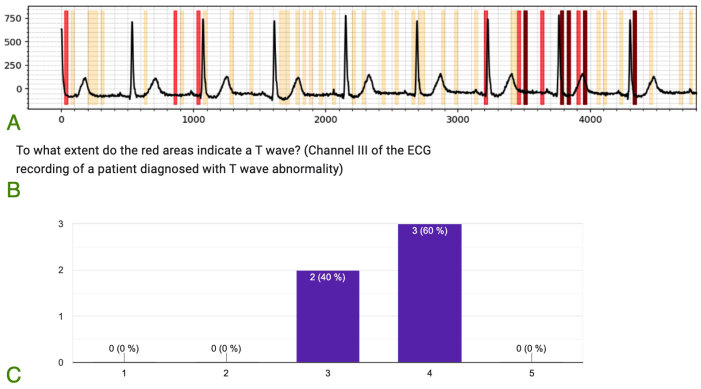


Figure 7. Example of interaction during the user study: respondents were presented with: (**A**) the most important lead, along with the ECG signal and explanations by means of heatmap coloring; (**B**) a particular question has been stated in connection with the selected diagnosis; (**C**) a summary of the answers of all five respondents to this question, is expressed as a bar plot.

Moreover, the respondents were allowed the option of expressing their reflections on the whole interaction by means of open text. In their text responses, the participants overwhelmingly felt that adding explanations in the presented form increased their confidence in these algorithms, confirming that machine learning algorithms need explainability. However, several respondents expressed distrust in algorithms that use only ECG recordings for prediction, as an electrophysiological examination consists of ECG, HRA (High Right Atrium), HBE (His Bundle Electrogram), and RVA (Right Ventricular Apex) recordings. In addition, doctors use other data,

such as patient records, echocardiographic examinations, and more. Limiting the data to only ECG records can cause experts to distrust prediction algorithms.

Finally, we discuss a number of issues with perturbation-based approaches and their implications for our method. Our method uses the perturbation principle at two levels (selection of the most relevant channel and selection of the most important parts of the ECG signal within that channel), and the issues discussed below can be influential at both of them. In general, perturbation-based methods select features that are important to the model; these are not necessarily the ones with the highest intrinsic explanatory power, as they are based on the model's behavior and not on the underlying structure of the data. As a result, there might be other intervals of the ECG signal that are more relevant, even for more sophisticated models. Further testing of the proposed perturbation-based method should be performed using more sophisticated models in order to verify the robustness of our method.

This issue is less critical for the first stage, where the most significant ECG channel is selected, because the decision there is based on a set of examples from a particular class, with the most frequently voted channel being selected. Moreover, there is always the possibility of visualizing more channels with significant numbers of votes.

Another drawback of the perturbation-based methods is that it can devalue correlated features. When two or more features are highly correlated, perturbing or removing one feature can result in similar changes in the model's output. This can lead to an underestimation of the importance of correlated features. This effect can be worse for more complex models, as they can learn more sophisticated relationships between features and rely more heavily on correlated features. One way to address this problem is to use different kinds of perturbations. In our method, we optimized only one parameter, the length of the perturbed interval. Another possibility could be to combine our approach with other interpretation methods.

Our experimental design presents human experts with the model's output, then asks them about the extent to which they agree. This may cause an anchoring bias, where the human expert's default response might be to agree with the model and to only disagree in extreme situations. To prevent this potential bias, experts could be asked to first label the important regions based on their opinion without showing them the model's predictions, then compare the selections of the expert with those of the model. This would require a completely different experimental design, however, which could be a possibility for future research.

6. Conclusions

Our analysis of the state-of-the-art showed that black-box models are among the most popular and successful models in the field of ECG data analysis. At the same time, the area of sensory data turned out to be the least researched. The lack of freely available libraries has caused many authors to transform sensory data into images, then treat the sensory data as image data. Therefore, we focused on processing the original sensory data from the ECG. However, the actual data visualization for doctors is provided as an image in a form they are accustomed to. We designed, implemented, and verified a model based on a convolutional neural network. Subsequently, we implemented our own perturbation method for explainability and verified the results through a user study.

The explainability method was implemented in Python, and its results can be displayed graphically in the form of heatmap coloring projected on the ECG graph. It contains two main functions. First, it selects and displays the most important data stream, followed by the most important areas within it for specific diagnoses. Verification was based on close communication with experts during the preparation and subsequent implementation of the user study. The function for plotting the most important streams contained very good explainability. The feature showing the most important areas on a particular ECG lead was able to correctly mark only a part of the relevant lead areas.

Available public databases with ECG recordings have allowed us to advance research in the area of black-box models and their explainability in the context of ECG signal data.

However, these models are distrusted by experts due to their limited viewpoint, as they only use ECG recordings. Therefore, from the point of view of further research, studying the combination of neural networks for multi-modal data appears to be a suitable direction.

Author Contributions: Conceptualization, J.P. and M.K.; methodology, M.K., J.P. and A.J.; software implementation, A.J.; validation, A.J. and Z.P.; expert consultations, Z.P.; resources, A.J.; writing—original draft preparation, J.P., M.K. and O.L.; writing—review and editing, all; visualization, A.J. and M.K.; supervision, J.P.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Slovak Research and Development Agency under contract No. APVV-17-0550 and contract No. APVV-20-0232, and by the Scientific Grant Agency of the Ministry of Education, Science, Research, and Sport of the Slovak Republic and the Slovak Academy of Sciences under grant number 1/0685/21.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rojat, T.; Puget, R.; Filliat, D.; Del Ser, J.; Gelin, R.; Díaz-Rodríguez, N. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv* **2021**, arXiv:2104.00950.
2. Šimić, I.; Sabol, V.; Veas, E. XAI Methods for Neural Time Series Classification: A Brief Review. *arXiv* **2021**, arXiv:2108.08009.
3. Vijayarangan, S.; Murugesan, B.; Vignesh, R.; Preejith, S.; Joseph, J.; Sivaprakasam, M. Interpreting deep neural networks for single-lead ECG arrhythmia classification. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 300–303.
4. Strodthoff, N.; Strodthoff, C. Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiol. Meas.* **2019**, *40*, 015001. [CrossRef] [PubMed]
5. Raza, A.; Tran, K.P.; Koehl, L.; Li, S. Designing ecg monitoring healthcare system with federated transfer learning and explainable ai. *Knowl.-Based Syst.* **2022**, *236*, 107763. [CrossRef]
6. Jeyakumar, J.V.; Noor, J.; Cheng, Y.H.; Garcia, L.; Srivastava, M. How can i explain this to you? An empirical study of deep neural network explanation methods. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4211–4222.
7. Ivanovs, M.; Kadikis, R.; Ozols, K. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognit. Lett.* **2021**, *150*, 228–234. [CrossRef]
8. Schlegel, U.; Arnout, H.; El-Assady, M.; Oelke, D.; Keim, D.A. Towards a rigorous evaluation of xai methods on time series. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 4197–4201.
9. Reyna, M.A.; Alday, E.A.P.; Gu, A.; Liu, C.; Seyedi, S.; Rad, A.B.; Elola, A.; Li, Q.; Sharma, A.; Clifford, G.D. Classification of 12-lead ecgs: The physionet/computing in cardiology challenge 2020. In Proceedings of the 2020 Computing in Cardiology, Rimini, Italy, 13–16 September 2020; pp. 1–4.
10. Singstad, B.J.; Tronstad, C. Convolutional Neural Network and Rule-Based Algorithms for Classifying 12-lead ECGs. In Proceedings of the 2020 Computing in Cardiology, Rimini, Italy, 13–16 September 2020; pp. 1–4. [CrossRef]
11. Nelder, J.A.; Mead, R. A simplex method for function minimization. *Comput. J.* **1965**, *7*, 308–313. [CrossRef]
12. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
13. Islam, S.R.; Eberle, W. Implications of Combining Domain Knowledge in Explainable Artificial Intelligence. In Proceedings of the AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering, Palo Alto, CA, USA, 22–24 March 2021.
14. Blahút, P. ECG & Arrhythmology (a Book Written in Slovak). 2019. Available online: https://www.techmed.sk/ekg-a-arytmologia-kniha/ (accessed on 7 May 2022).
15. Suresh, H.; Hunt, N.; Johnson, A.; Celi, L.A.; Szolovits, P.; Ghassemi, M. Clinical intervention prediction and understanding using deep networks. *arXiv* **2017**, arXiv:1705.08498.