Institute of Computer Science at Freie Universität Berlin

BACHELOR THESIS

Generating Counterfactual Explanations for Electrocardiography Classification with Native Guide

Viktoria Andres

Artificial Intelligence and Machine Learning Work Group

First Examiner:	Prof. Dr. Eirini Ntoutsi
Second Examiner:	Prof. Dr. Claudia Müller-Birn
Supervisor:	Philip Naumann

September 24, 2021

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. The thesis was not examined before, nor has it been published. The submitted electronic version of the thesis matches the printed version.

Berlin, September 24, 2021

Viktoria Andres

Abstract

Explanations are essential components in the promising fields of artificial intelligence (AI) and machine learning. Deep learning approaches are rising due to their supremacy in terms of accuracy when trained with huge amounts of data. Because of their black-box nature, the predictions are also hard to comprehend, retrace, and trust. Good explanation techniques can help to understand why a system produces a certain prediction and therefore increase trust in the model. Understanding the model is crucial for domains like healthcare, where decisions ultimately affect human life. Studies have shown that counterfactual explanations in particular tend to be more informative and psychologically effective than other methods.

This work focuses on a novel instance-based technique called "Native Guide", that generates counterfactual explanations for time series data classification. It uses nearest neighbour samples from the real data distribution with class change as a foundation. This thesis applies the method on the explanation of electrocardiogram (ECG) classification, a very complex and vital medical field where every single ECG carries unique features. Native Guide for ECGs is explained, examined and expanded by providing necessary background knowledge, amplifying aspects like plausibility, comparing different suitable models to each other and indicating benefits and downsides. Finally, counterfactual explanations for ECG data classification generated by Native Guide are evaluated by cardiologists by means of two expert interviews.

Synchronization of the periodic ECG data was shown to be the most important contribution to the method that enabled the generation of plausible counterfactuals. The experts, who had never seen or used counterfactuals in their work, were interested in this approach and could envision its application within the field when it comes to training junior doctors. In general, AI classification along with sophisticated proximate counterfactuals indicate success and reliability when it comes to the identification of heart diseases.

Zusammenfassung

Erklärungen sind wesentliche Komponenten in den vielversprechenden Bereichen der KI und des maschinellen Lernens. Deep-Learning-Ansätze sind auf dem Vormarsch, weil sie beim Training mit riesigen Datenmengen eine präzise Genauigkeit aufweisen. Aufgrund ihrer Blackbox-Natur sind die Vorhersagen aber auch äußerst schwer zu verstehen, nachzuvollziehen und zu vertrauen. Gute Erklärungstechniken können helfen zu verstehen, warum ein System eine bestimmte Vorhersage macht, und somit das Vertrauen in das Modell steigern. Das Verständnis des Modells ist in Bereichen wie der Gesundheitsfürsorge, wo Entscheidungen letztlich das Leben von Menschen beeinflussen, von entscheidender Bedeutung. Studien haben gezeigt, dass insbesondere kontrafaktische Erklärungen informativer und psychologisch wirksamer sind als andere Methoden.

Diese Arbeit befasst sich mit einer neuartigen instanzbasierten Technik namens „Native Guide“, die kontrafaktische Erklärungen für die Klassifizierung von Zeitreihendaten generiert. Sie verwendet Stichproben aus der realen Datenverteilung mit Klassenwechsel als Grundlage. Die Arbeit wendet die Methode auf die Erklärung der Klassifizierung von Elektrokardiogrammen (EKG) an, ein sehr komplexes und wichtiges medizinisches Gebiet, in dem jedes einzelne EKG einzigartige Merkmale aufweist. Native Guide für EKGs wird erklärt, untersucht und erweitert, indem notwendiges Hintergrundwissen vermittelt wird, Aspekte wie Plausibilität verstärkt, verschiedene geeignete Modelle miteinander verglichen und Vor- und Nachteile aufgezeigt werden. Schließlich werden kontrafaktische Erklärungen für die Klassifizierung von EKG-Daten, die mit Native Guide generiert wurden, von Kardiologen im Rahmen von zwei Experteninterviews bewertet.

Es stellte sich heraus, dass die Synchronisation der periodischen EKG-Daten der wichtigste Beitrag zur Methode war und die Generierung plausibler kontrafaktischer Erklärungen ermöglichte. Die Experten, die in ihrer Arbeit noch nie Kontrafakten begegnet sind, zeigten sich an dem Ansatz interessiert und konnten sich verschiedene wichtige Anwendungsfelder vorstellen, wie z.B. die Ausbildung von Assistenzärzten in der Inneren Medizin. Im Allgemeinen versprechen KI-Klassifizierung und ausgereifte, näher liegende Kontrafakten eine erfolgreiche Zukunft für eine zuverlässigere Erkennung von Herzkrankheiten.

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

Over the last few months, I dedicated myself to tackle the research problem of explaining the often called unexplainable deep learning models through counterfactuals in the incredibly interesting and crucial but also deeply complex field of ECG classification. To accomplish this objective, I needed to be a generalist as well as a specialist. As a generalist I needed to know all general background knowledge on the topics, and as a specialist I needed to know the intricacies of electrocardiography and be able to explain counterfactuals within the Native Guide. I am thankful for all the opportunities to learn and develop my skills and knowledge in the field of XAI during the different stages of writing this thesis.

Special thanks goes to all people that supported me. First of all, I want to thank my supervisors and advisors that guided and helped me with all my questions and problems during the whole time. Second, thanks to the cardiologists that made it possible to gain deep insights into day-to-day work with ECGs and prospectively counterfactuals. Third, I want to thank all my friends that provided me with valuable discussions, reviews and helped me to keep focus on the work and progress.

Viktoria Andres

1. Introduction

1.1. Topic and Related Work

The field of artificial intelligence (AI) continues to evolve and become more and more complex. Many machine learning algorithms, especially those that include artificial neural networks, achieve exceptional performance and accuracy. To give an example, DeepMind's AI AlphaFold determines a protein's 3D shape from its amino-acid sequence and therefore solved one of biology's greatest challenges [1]. AlphaGo, also an AI of DeepMind, defeats international champions in the boardgame Go [2]. A serious problem with accurate but deeply complex systems is interpretability. Greater attention for explainable AI has recently become important with new research into modern deep learning models and their black-box character [3]. More research is needed to realize sufficiently explainable models in all the different systems, data and use cases. This work focuses on time series data, specifically electrocardiograms (ECG), and their classification. Classification tasks predict class memberships for specified data inputs [4]. For instance, predicting whether an electrocardiogram corresponds to a normal or abnormal class. "Native Guide" is a method to generate instance-based counterfactual explanations on time series data [5]. Counterfactuals are a subclass of different explanation techniques, that try to show alternatives to key elements of the data with a different outcome [6][7][8].

Besides Native Guide there are many approaches that try to generate good counterfactual explanations [5][9]. However, only few can be applied directly to the time series classification field [5][10][11]. Counterfactuals for time series data is still a largely unstudied problem. Methods from other fields, like image classification, cannot be directly adjusted to time series models due to the very complex nature of time series [12]. Quite similar to Native Guide, E. Ates et al. [10] described a counterfactual generation approach for multivariate time series data using a greedy search algorithm on distractor instances. I. Karlsson et al. [11] explained a method for univariate time series tweaking, that does not require instances of the training data. A. Gee et al. [13] introduced prototypes as maximal representatives of a class, but do not make the connection to counterfactuals. Their work helps to get global insights into time series classification. Still, it is less useful looking at more discriminative areas of the time series where samples can have unique features [5]. When it comes to using counterfactuals or other explanation methods, user studies like in the papers [14],[15] and [16] have shown that users preferred counterfactuals more than other explainable methods like case-base reasoning or the feature importance approach. Thereby, this thesis focuses on the counterfactual ex-

plainability method Native Guide. In the original paper [5] by E. Delaney et al. the method was promised to generate plausible, proximate, diverse and sparse counterfactuals by design.

1.2. Research Approach: Design Science

A Design Science [17] project centers around an artifact as the object of study. To thoroughly investigate the artifact two key components of the project, the design and the investigation, must be carefully constructed and followed. An artifact interacts with a problem context to improve that context in any possible way. Therefore, Design Science problems are also referred to as improvement problems. Designs are solutions to a problem and artifacts aim for improvements in the researcher's design. Methods, algorithms, or conceptual structures that hold all requirements are artifacts [17][18].
This thesis investigates the problem context of counterfactual explanations in time series data. It also investigates the Native Guide method as the design artifact using existing problem-solving knowledge and newly acquired knowledge from the investigative and observational steps.

1.2.1. Aims and Objectives

The goal of this thesis is to investigate different aspects about Native Guide, a generative counterfactual explanation method [5]. First, an in-depth overview of background information is provided about the proposed method of inquiry. Second, we investigate the method and compare three different classifiers for time series data (Fully Convolutional Neural Network [19], Residual Network [19] and InceptionTime [20]). Third, we go one step further and extend the method with additional ideas involving different altering of the generated counterfactual through synchronization, another perturbation approach and a shifted decision boundary. This should enable more reasonable use in the explanation of complex ECG data classification. Fourth, expert interviews evaluate the counterfactuals from a user's perspective and share information about the problem space. And finally, critical examination of the benefits and drawbacks of the approach are examined.

1.2.2. Research Questions

All following research questions (RQ) should be interpreted in the context of time series ECG classification and counterfactual explanation using the later explained Native Guide method:

RQ1 Which scientific knowledge is needed to understand and implement the Native Guide method?

RQ2 Are there further technical approaches that could enrich the counterfactual goodness but still maintain the basic idea of the method?

RQ3 Does a counterfactual build trust in the prediction of a model?

RQ4 Does a counterfactual provide more insights about data than a prediction without counterfactual?

1.3. Thesis Outline and Research Methods

To answer the above stated research questions (RQ), various research methods were applied. We systematically reviewed the knowledge needed to design and develop the Native Guide method through literature research, answering RQ1 in the chapter *2 Background*. Next, an experimental approach using observations, qualitative and quantitative investigations and optimizations examines aspects about the method and aims to improve them. This information provided answers for the extensive question RQ2 in chapter *3 Native Guide: A Counterfactual Explanation Technique*. Further, we evaluate the method in a qualitative problem-centered expert interview on complex ECG data with cardiologists. Answers to research questions RQ3 and RQ4 in chapter *4 Evaluation: Expert Interview*, were obtained through expert interviews which provided more insights about the current use of ECGs. These interviews also provided insight for the future of predictive AI models and counterfactuals in the medical cardiology field. Lastly, the chapters *5 Discussion* and *6 Conclusion and Future Work* complete the thesis.

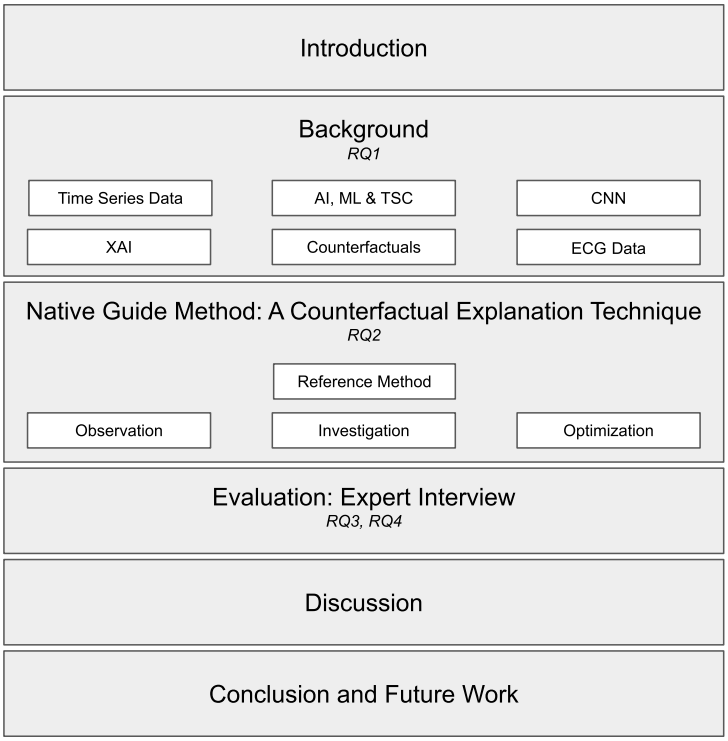


Figure 1.1.: Thesis Overview

2. Background

The next sections provide an overview of the required background knowledge for a counterfactual explainability method in the field of ECG classification. First, we introduce time series data in chapter 2.1. Second, we want to share a basic understanding of artificial intelligence, machine learning, and classification in 2.2. In behalf of that, we also introduce convolutional neural networks (CNN) more specifically in 2.3. Next, we explain explainable artificial intelligence (XAI) in 2.4 and counterfactual explanations in 2.5. Lastly, we offer insights into ECG signal data and corresponding datasets in the sections 2.6 and 2.7.

2.1. Time Series Data

Time series are collections of observations obtained through repeated measurements over time [21]. Any data with an ordered set of values can be adapted to a time series [22][23]. The following are formal defintions:

Definition 1 (Time Series) *A time series* $T = \{t^{(1)}, t^{(2)}, \cdots, t^{(n)}\}$ *is a vector, where* n *is the length of the time series and each point* $t^{(k)} = (t_1^{(k)}, t_2^{(k)}, \cdots, t_m^{(k)}) \in \mathbb{R}^m$ *is an array of* m *values [5][22].*

Definition 2 (Time Series Data Set) *A time series dataset* $D = \{T_1, T_2, \cdots, T_q\}$ *is a collection of time series, where each time series* $T_k \in \mathbb{R}^{n \times m}$ *may have a* ***class label*** $y_k \in \mathbb{Z}$ *from a set of class labels* $Y = (y_1, y_2, \cdots, y_c)$ *[5].*

Definition 3 (Univariate) *A univariate time series* T_U *is a time series, where each point* $t^{(k)} \in \mathbb{R}^m$ *is a 1-dimensional vector with* $m = 1$ *[24].*

Definition 4 (Multivariate) *A multivariate time series* T_M *is a time series, where each point* $t^{(k)} \in \mathbb{R}^m$ *is a m-dimensional vector with* $m > 1$ *[24].*

Despite its omnipresent character, time series data analysis (TSA) is less researched than the analysis of other types of data, such as images or tabular data, especially for sub-domains. Different characteristics and temporal interdependencies make time series data difficult to analyse and predict future behaviour [22].

To give an overview, we define some global characteristics of time series data:

- **Trend** in a time series means that there is a long-term change in its mean level. In other words, the time series center around an increasing or decreasing line which suggest a trend [25].
- **Seasonality** patterns repeat themselves regularly and predictable over intervals of time with fixed length, often on a daily, weekly or monthly occasion [25].
- **Periodicity** is a cyclical pattern that (like seasonality) also repeats itself over time, but varies in frequency or cycle length [25].
- **Serial correlation** is the observed interdependence between a variable and its subsequent in the time series over periods of time [25].
- **Skewness** is the lack of symmetry left and right of the center point in a distribution [25].
- **Kurtosis** describes a peaked or flat characteristic, relative to a normal distribution. High kurtosis tend to have distinct peaks near the mean, rapid drops and heavy tails. Low kurtosis usually have a flat top near the mean and less sharp peaks [25].
- **Linearity** is a property of a time series, where each data point is a linear combination of past or future data points. Nonlinear time series usually describe much more complex dynamics [25][26].
- **Self-Similarity** can be observed, if a time series has a substructure very similar or identical to its overall structure. This implies long-range dependencies in the time series [25][27].
- **Chaos** is a random processes with sensitive dependence on initial values, which is also described as the butterfly effect. In a nutshell, very small changes or errors of a value lead to very big changes in the near future. Similar causes have completely different effects in chaotic time series [25][28].

2.2. Basic Understanding of Artificial Intelligence, Machine Learning, and Classification

Artificial intelligence (AI) has gained different definitions over the past several years. We define AI corresponding to D. Poole et al. as "a field that studies the synthesis and analysis of computational agents intelligibly" [29].

Machine learning is a subfield of artificial intelligence and defined as a collection of computational methods that learn a model with data to improve

performance or to make predictions [30]. Such a model f takes an input data sample X and produces an output class label Y. We learn the model, so that an input space gets transformed into a correctly mapped output space $f : X \rightarrow Y$.

Deep learning is a subfield of machine learning that involves learning models with a high level of abstraction. They are composed of multiple processing layers, which create models that are referred to as artificial neural networks (ANN), that perform nonlinear transformations on their input. To work efficiently, a network needs large amount of representative data. Deep learning is called deep, because of several layers of neurons that are stacked one after another [31][32]. Artificial neurons are analogous to biological neurons. They are the fundamental component for building artificial neural networks. Neurons receive inputs and produce outputs applying different parameters. This unit is also called a perceptron. Neural networks consist of multilayer perceptrons (MLP) which contain one or more hidden layers with multiple hidden neurons in them [33]. Figure 2.1 shows a schematic representation of a neuron and an ANN compared to its human counterpart.

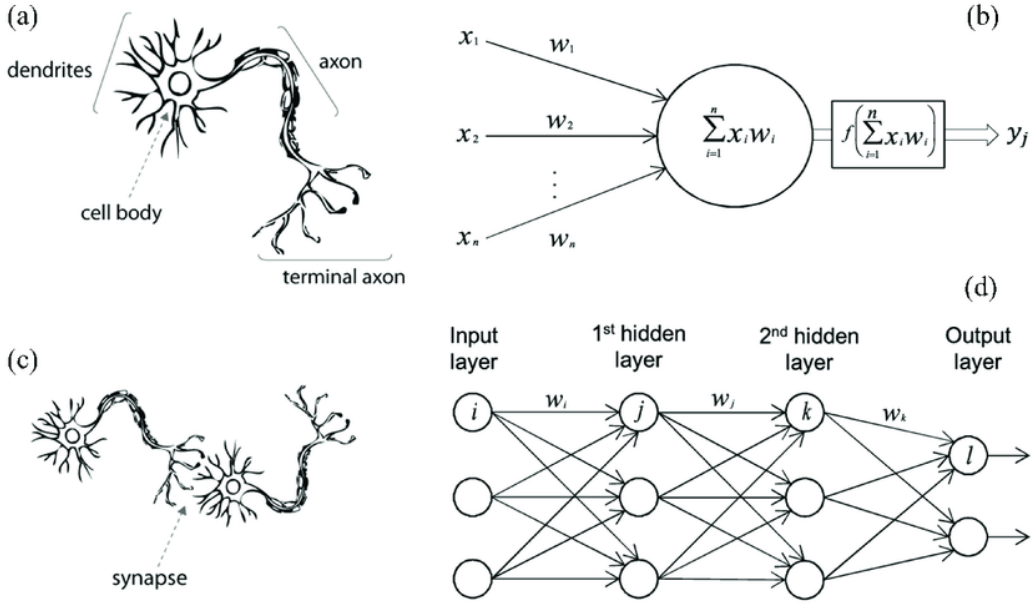


Figure 2.1.: Neurons and Artificial Neural Networks: (a) Human Neuron, (b) Artificial Neuron, (c) Biological Synapse, and (d) ANN Synapses [34]

A classification task in the field of machine learning is a problem that requires predicting class memberships for data instances. In this case, the output Y is a probability space where a model f assigns a specific class label from Y with highest probability to the model input X [4].

2.3. Convolutional Neural Network (CNN)

This thesis focuses on convolutional neural networks (CNNs) as a deep learning classification technique. The network structure is not a novel idea. It was first proposed by Fukushima [35] in 1988. However, computational limits made its application difficult at that time. Further improvement and development led to state-of-the-art results in many recognition tasks [33]. Popularity spread when AlexNet [36] had won the ImageNet competition in 2012 and CNN became an important tool for many domains [23].

We define the architecture of a CNN as follows: an input layer, convolutional layers, pooling layers, a feature or flattened data layer, and an output layer:

- **Input layer** takes a time series as an input and has $n \times m$ neurons, where m is the input time series dimension and n the length of each series [33].
- **Convolutional layer** applies and slides learnable filters (i.e. convolutions) over the time series of the preceding layer. A filter can be pictured as a generic non-linear transformation that averages specified proportions of a time series with a moving window. The output of the filters can go through a batch normalization function, that maintains the mean output close to zero and the output standard deviation close to one, and an activation function, such as the sigmoid or ReLu function, to form the so called output feature maps [23][33].
- **Pooling layer** divides a feature map into equal-length segments. In case of average of max pooling, every segment is represented by its average or maximum value. Pooling is down-sampling the feature maps and thus reducing variability in the hidden activations. However, not every CNN architecture incorporates pooling layers [33]. Global Average Pooling (GAP) does not divide the feature maps providing only one averaged segment for every feature that result in a single real value that represents a whole feature. A GAP layer can be used instead of a feature layer, as it is demonstrated in figure 2.2 [5].
- **Feature layer** (or flattened data layer) represents the original time series as a series of feature maps. Connecting all feature maps generates a new time series as the final representation of the original input in the feature layer [33].
- **Output layer** consists of c neurons that all correspond to a class in Y of a time series. It is fully connected to the feature layer. Then, usually the maximum output neuron represents the class label solving a classification task [33].

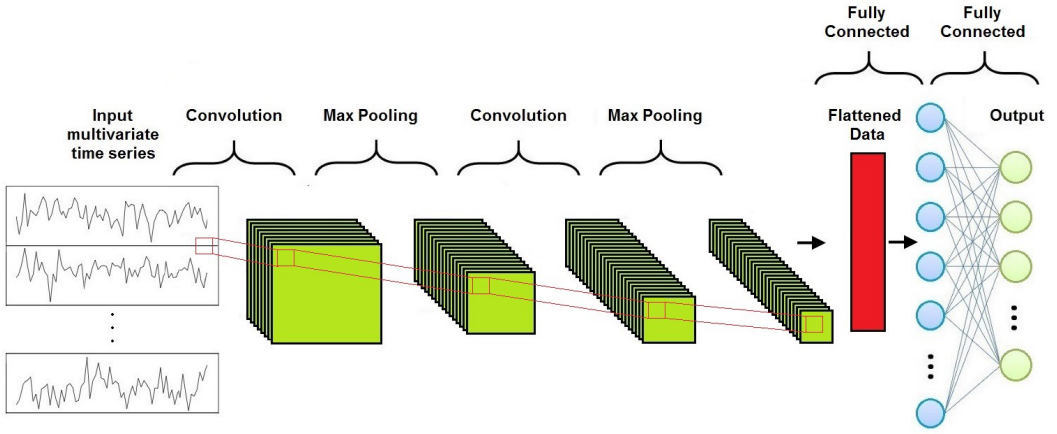


Figure 2.2.: CNN for Multivariate Time Series [37]

2.4. Explainable Artificial Intelligence (XAI)

In recent years, machine learning models have become more and more effective due to improved algorithms, ever growing datasets and exceeding computational power. Improved predictive accuracy also involves an increase in the complexity of the models. Deep learning systems have boosted the predictive power but have also decreased the ability to understand and interpret decisions of a model. The black-box character and complexity of the deep learning systems makes its inner workings hard to explain. On the other hand, white-box models typically produce explainable results, but lack in performance compared to the former [38][39].

A system, that does not provide a rationale for a model's prediction, cannot be fully trusted. To be able to validate a model's correctness, detect bias or leakages, and potentially learn new patterns in the data, a basic understanding of the model is crucial. This is vital for domains such as healthcare where decisions ultimately affect human life. There is a need for trustworthy, fair, and high-performing models that can explain their predictions and actions to their users. This is where explainable artificial intelligence comes into play [38][40][41]. Explainable artificial intelligence (XAI) is one of many human-agent interaction problems [8] that tries to combine excellent performance with a human understanding of predictions and psychologically effective explanations [41]. Human-agent interaction can be defined as the intersection of artificial intelligence, social science, and human-computer interaction (HCI) [8], as shown in figure 2.3.

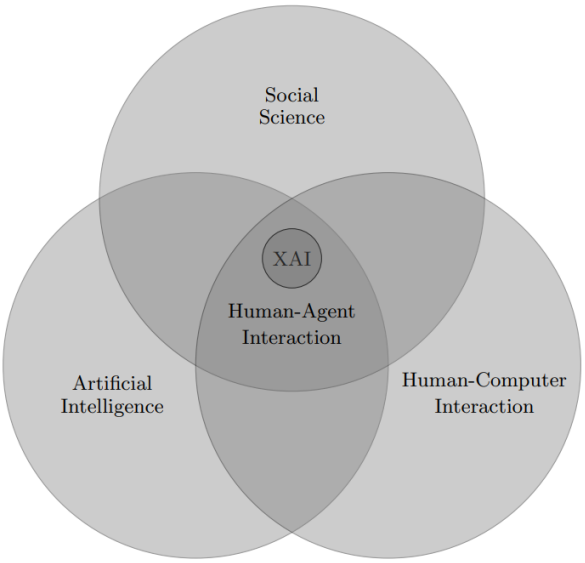


Figure 2.3.: Human-Agent Interaction [8]

A brief taxonomy of basic XAI method characteristics [38]:

- **Model-specific** methods are restricted to a related group of machine learning classifiers.
- **Model-agnostic** methods are applicable for any possibly classifier.
- **Local** methods produce explanations for one specific data instance.
- **Global** methods produce explanation for the whole model.
- **Post-Hoc** methods focus on analysing and explaining a black-box model after training by applying particular interpretation methods.
- **Intrinsic** methods restrict the model complexity trying to make it more interpretable due to a simpler structure.
- **Enhancing fairness** is another interpretability method that tries to fight inequalities and discrimination on model predictions. Ideally, machine learning algorithms should be impartial and non-discriminative. Different techniques try to remove bias from data and predictions and to train a model to make fair predictions.
- **Sensitivity analysis** is a method measuring the contribution of a single or multiple input variables to the output variance. That way the impact of each feature in the model prediction is examined.

2.5. Counterfactual Explanations

What might our life be like if we had made key choices differently? What if we had moved to another city, attended a different university or chose to have no kids? It is common to ask questions like that occasionally. In fact, these types of questions are counterfactuals [6].

We will now take a look at the logical understanding and intuition behind counterfactuals. Factual conditionals state that if one fact is true, then so is another ($p \implies q$) [42]. A factual condition in natural language would be:

> "If I wash my hands for 30 sec, then they get clean."

When the condition is known for sure it becomes a fact, and we can construct a counterfactual:

> "If I had not washed my hands for 30 sec, they would not have got clean."

That is just one possible counterfactual. Another variation is:

> "If I had not washed my hands for 30 sec, they would have got clean."

The next variation is not possible, because it contradicts to the initial assumption, and therefore it is not a valid counterfactual:

> "If I had washed my hands for 30 sec, they would not have got clean."

Finally, we could also alter the "30 sec" and get infinite different counterfactuals like:

> "If I had washed my hands for 10 sec, they would have got clean."

There are various use-cases for counterfactuals. One compelling use-case is in the field of XAI. Counterfactuals can help to make complex and incomprehensible black-box system predictions more intelligible to developers and users, thus paving the way for interpretable models. An interpretable model gives insights about aspects of the system and helps to understand the causes of its decision making process. It enables the user to see different (counterfactual) explanations, such as why one decision was made instead of another, and to predict how a change to a feature will affect the system's output [7][8].

For the scope of our black-box (time series) classification tasks, we define more specific and restricted counterfactuals as follows:

Definition 5 (Counterfactual) *Given an input $x \in X$ and the corresponding output class $c \in Y$ predicted by a classifier model $b(x) = c$, a counterfactual x' is a perturbed instance of x that generates a different output class $b(x') = c' \in Y$ [43].*

2.6. ECG Signal Data

This thesis will focus on time series data generated by an electrocardiogram (ECG). ECGs are used in the medical field to detect heart anomalies through activity measurements of electrical impulses produced by the heart. An ECG machine records this electrical activity and displays it as a trace on paper, which then is interpreted by a medical or cardiac expert. Damage or disorders of the heart can change its electrical activity and thus the ECG trace. ECG has become the most important tool when it comes to the diagnosis of heart diseases [44].

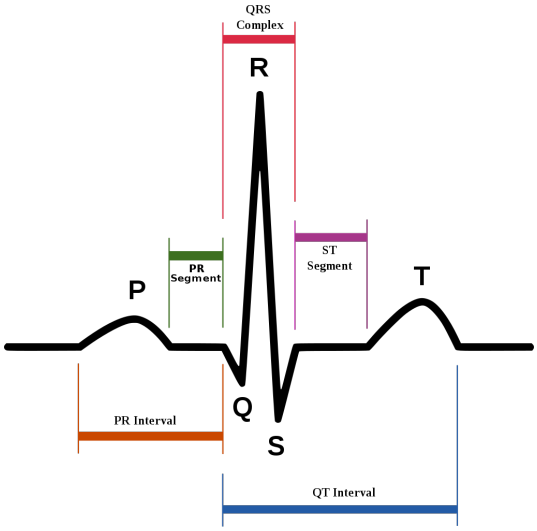


Figure 2.4.: ECG Waveform [45]

ECG signals are periodic time series where each beat consists of a P wave, QRS complex, and a T wave. We distinguish between different peaks (P, Q, R, S, and T), intervals (PR, QRS, ST, and QT), and segments (PR, ST), all with inherent characteristics like amplitude and duration [45], which can be seen in figure 2.4.

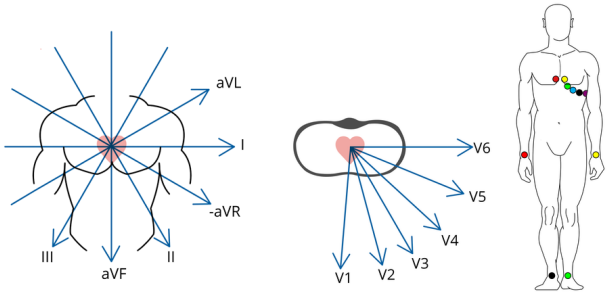


Figure 2.5.: Lead Positions and Axes [46]

There are different approaches to produce an ECG. A standard ECG has 12 leads, which measure the heart activity. A lead is a pair of electrodes (+ve and -ve) placed on different regions of the body and connected to an ECG recorder, as shown in figure 2.5. Bipolar leads record the potential difference between two points. Unipolar leads, on the other hand, record the electrical potential at one particular location by a single electrode [45].

Characteristics	Lead	Positions
Einthoven, Bipolar Leads	I	Potential between left and right arm
	II	Potential between right arm and left leg
	III	Potential between left arm and left leg
Limb, Unipolar Leads	aVR	+ve on right arm
	aVL	+ve on left arm
	aVF	+ve on left leg
Chest, Unipolar Leads	V1	4th intercostal space, right sternal edge
	V2	4th intercostal space, left sternal edge
	V3	between 2nd and 4th electrodes
	V4	5th intercostal space in midclavicular line
	V5	5th rib, anterior axillary line
	V6	midaxillary line

Table 2.1.: Leads Description [45]

Classification of ECGs is very complex because different abnormalities do look different and can be very subtle. They may occur only in specific leads. That is why it is essential to produce multi-lead ECGs. Nevertheless, figure 2.6 shows one example of a normal and an abnormal ECG signal, specifically an abnormal change of the ST segment and T peak [47].

Figure 2.6.: Normal vs. Abnormal ECG [47]

Early detection and discrimination between different disorders is crucial for proper treatment of patients. Classification using machine learning algorithms can help cardiologists to identify and distinguish heart problems and save lifes by providing the proper therapy [44].

2.7. Openly-accessible ECG Datasets

There are not many ECG datasets openly accessible that are also sufficient to train an artificial neural network. This thesis examined four different data collections with different qualities, presented in the following.

2.7.1. ECG200

The ECG200 dataset has 200 ECG samples each with only one heartbeat waveform and 96 data points. It is a binary 1-lead univariate dataset. The dataset can be easily accessed through the UCR time series archive [48][49], but the small amount of data makes is almost impossible to learn a good deep neural network with much more learnable parameters than provided data samples [50].

Class ID	Class Name	Amount
0	Normal	133
1	Myocardial Infarction	67

Table 2.2.: ECG200

2.7.2. ECG5000

The ECG5000 dataset is also accessible through the UCR archive [48][49] and holds 5000 samples extracted from a single patient and is divided into five different classes. Each data sample is 140 points long [51]. Even though it has more samples than ECG200, it still does not provide enough to learn a good deep neural network [50], especially for class 2, 3, and 4.

Class ID	Class Name	Amount
0	Normal	2927
1	R-on-T Premature Ventricular Contraction	1767
2	Premature Ventricular Contraction	96
3	Supraventricular Premature Beat	194
4	Unclassified Beat	24

Table 2.3.: ECG5000

2.7.3. PTB

A ECG heartbeat categorization dataset on kaggle [52] used the data from the PTB diagnostic ECG database [53] to prepocess it to enable useful classification with deep learning algorithms. Their preprocessed dataset consist of 14552 samples and is using only one lead (univariate) instead of the original 15 leads in PTB. The sampling frequency is 125 Hz resulting in 188 data points as the

length of an ECG. It is a binary 1-lead univariate dataset differentiating between normal and abnormal ECGs. Classification examples seem to accomplish good accuracy. However, to obtain equal length shorter samples are padded with zeros in the end and peaks are not synchronized, which makes it harder to use such data for the Native Guide method.

Class ID	Class Name	Amount
0	Normal	4046
1	Abnormal	10506

Table 2.4.: PTB

2.7.4. PTB-XL

The PTB-XL dataset consist of 21837 samples collected from 18885 different patients. The ECGs have 12-leads, so the data is multivariate. It represents one of the largest freely accessible ECG dataset. The instances are collected over a time period of 10 seconds with a sampling frequency of 100 Hz. Therefore all data samples have 1000 data points per lead. Furthermore, it is a multi-label dataset labeled by two doctors, which means that one sample can be assigned to multiple classes. Being able to assign a sample to multiple classes more closely mirrors reality since patients can have more than one abnormality [47].

Class ID	Class Name	Amount	Amount (Preprocessed)
0	Conduction Disturbance	4907	4492
1	Hypertrophie	2655	2650
2	Myocardial Infarcion	5486	5485
3	Normal	9528	9083
4	ST/T Change	5250	5217

Table 2.5.: PTB-XL

All things considered, we work with the PTB-XL dataset for later investigation and optimization of the method. To make the data more reasonable some time series can not be considered for analysis. All time series that are labeled as normal together with any other abnormal class are illogical and thus excluded during preprocessing. We assume that such labels exist because of different classifications proposed by the doctors.

3. Native Guide: A Counterfactual Explanation Technique

The following chapters describe a novel counterfactual explanation technique called "Native Guide", originally proposed by E. Delaney et al. [5]. After explaining the method and implementation, further adjustments and extensions are introduced to reach more generalizable and proficient explanations. In the following we refer to the time series, of which we want to produce a counterfactual, the query time series T_q.

3.1. Reference Method

Unlike the original paper [5], this thesis defines Native Guide as an instance-based model-specific post-hoc explanation technique that generates counterfactual explanations for time series classifiers. The following points clarify the basic characteristics of the method and different views compared to E. Delaney et al. [5]. After that, we define and explain the four technical steps of the generation process based on [5].

- **Instance-based:** The generation process is based on existing instances in the dataset. This involves a given query time series that needs to be explained and the closest neighbour time series that is in another class. This way the generated time series are usually very similar to the query time series and in distribution of the data [5]. This also means that the method not only requires access to the classification model but also the training data, or at least a representative part of it.

- **Model-specific:** The paper [5] called this method model-agnostic, meaning it should be applicable for any possibly classifier. Given the fact that the algorithm requires feature maps and weights from the last model layers [5], we can argue that the method is model-specific. Not all ANNs incorporate these.

- **Generation:** The method generates new time series data that is useful to explain predictions of a model. This also has potential for data augmentation in sparse time series datasets [5].

3.1.1. Learn or Load Classifier

The first basic step of the method is to get an underlying (black-box) classifier b for the data that has to be explained. This classifier either needs to be

trained on respective data, that represents the classes for the to be explained data, or be pretrained on other but similar data and then fine-tuned to get better accuracy. It also needs to meet certain requirements that make the next steps possible. First, the model has to produce feature maps, which are the outputs of filters applied to the outputs of a previous layer [54]. Second, the model needs some kind of global-average-pooling (GAP) layer, where all feature maps get compressed to one value each. This precedes a dense layer, which associates the different feature values to all possible classes with a weighted connection [5][23].

Models without a GAP layer can not be used, since it is a key element for the next chapters 3.1.2 and 3.1.4. Known approaches, that produce feature maps and employ a GAP layer are: Fully Convolutional Neural Network (FCN), Residual Network (ResNet) and InceptionTime. All three are described in the following.

Fully Convolutional Neural Network (FCN)

FCN [55] for time series data was first proposed by Z. Wang et al. [19] and evaluated by H. I. Fawaz et al. [23] and N. Strodthoff et al. [47]. It is a convolutional neural network without any local pooling layers. The consequence is that the length of the time series are kept unchanged throughout the layer operations. The architecture replaces the final fully connected layer with a Global Average Pooling (GAP) layer reducing the number of parameters and enabling class activation mapping. As illustrated in figure 3.1, a FCN consists of three convolutional layer blocks. Each block contains three operations: a convolution, a batch normalization and a ReLU activation function in the end [19][23].

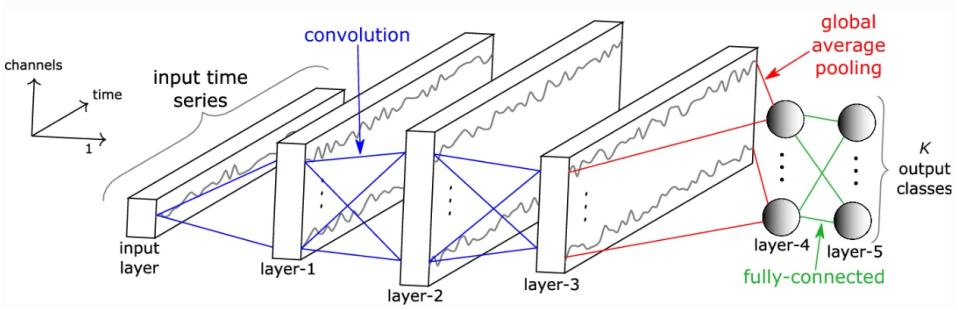


Figure 3.1.: FCN [23]

Residual Network (ResNet)

Likewise, ResNet [56] for time series was also first proposed by Z. Wang et al. [19] and evaluated by H. I. Fawaz et al. [23] and N. Strodthoff et al. [47].

The architecture is pictured in figure 3.2 and consists of 12 layers, of which two are the input and output layers, nine are convolutional layers and one is a GAP layer. Each convolution is followed by batch normalization and a ReLU activation function. ResNet also has a shortcut residual connection between each three consecutive convolutional layers, which also explains the name of the architecture. The linear shortcut enables flow of gradients directly through these connections, which makes training easier by reducing the vanishing gradient effect. This effect happens when the partial derivatives that control the training process are becoming smaller and smaller, which mitigates prediction performance [23].

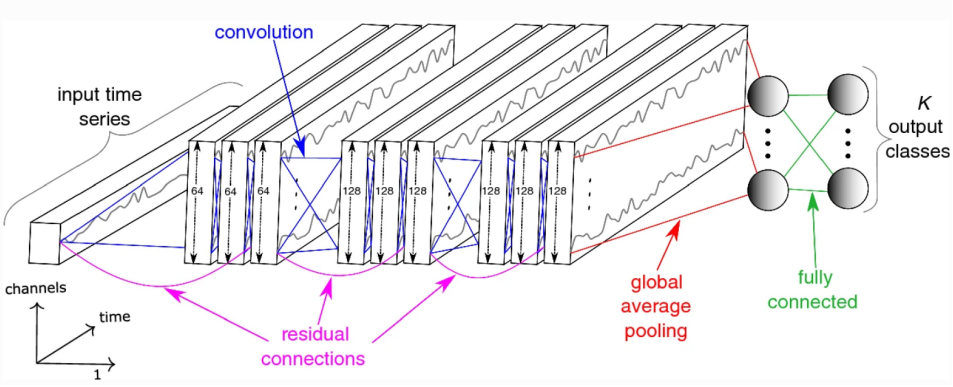


Figure 3.2.: ResNet [23]

InceptionTime

InceptionTime was introduced by H.I. Fawaz et al. [20] and compared to FCN and ResNet by N. Strodthoff et al. [47]. It is described as an ensemble of five Inception networks [57], with each prediction given an even weight. As shown in figure 3.3, InceptionTime consists of two blocks with each three inception modules rather than fully convolutional layers. Besides an input, output, and GAP layer, it also has two residual connections. It has one residual connection less than ResNet, which has three. Inception modules perform convolutions, but also incorporate bottleneck layers that reduce the dimensionality of multivariate time series and model's complexity. They also mitigate overfitting problems for small datasets, where the system memorizes the training data instead of learning predictive rules. Also, Inception modules allow to have much longer filters than ResNet with about the same number of parameters to be learned [20]. The convolution is simultaneously applying multiple filters of different lengths on the same input time series. Furthermore, parallel max pooling makes the model invariant to small perturbations.

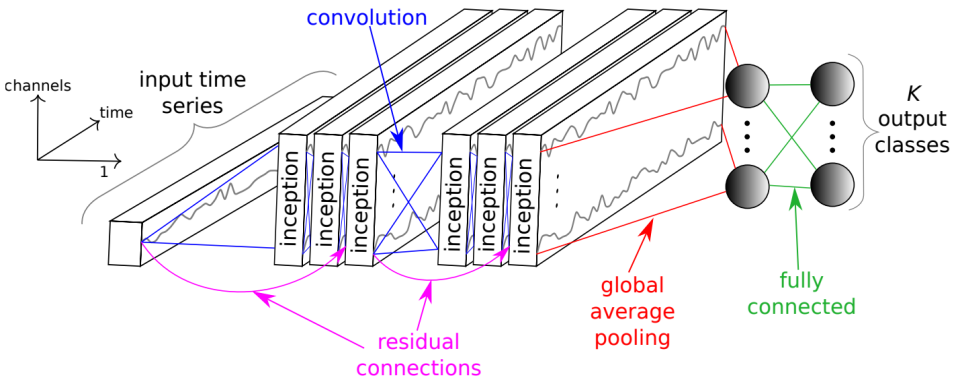


Figure 3.3.: Inception [20]

3.1.2. Class-Activation-Map (CAM)

The second step of the method produces a Class Activation Map. This process can be done for every time series in the dataset. First, the feature maps from the last convolutional layer are collected. One could picture feature maps for a time series input as an array with the length of that time series. Each element in that array is an importance value for every point in the specific input time series and corresponds to the features extracted by the applied filter. The higher the value the more important a data point is for that feature. Second, the weights that connect the global average pooling with the last dense layer are obtained. There should be weights from every feature value at the GAP layer to every class. The weights suggest how important a feature is to belong to a class. The higher the value, the more important is that feature for a class. Finally, the weighted sum of all feature maps is calculated for every time series. This process is done by multiplying every feature map with its weight for the class of the time series and adding the weighted feature maps together to get an importance value for each data point for that class [5]. We additionally illustrate the procedure in the algorithm 1.

Algorithm 1 Get CAMs for each T_q

Require: A classifier with feature maps, GAP and last dense layer

1: $cam \leftarrow \{\}$
2: **for each** T_q in dataset **do**
3: $\quad f \leftarrow getLastFeatureMaps(classifier, T_q)$
4: $\quad w \leftarrow getWeightsFromLastLayer(classifier)$
5: $\quad cam \leftarrow (w * f) + cam$
6: **end for**
7: **return** cam

getLastFeatureMaps() returns the feature maps from the last convolutional layer.
getWeightsFromLastLayer() returns the weights from the last dense layer.

3.1.3. Finding the Native Guide

The third step of the method is to search for a query's (T_q) counterfactual nearest (unlike) neighbour (T_{NUN}), also called "native guide". This is a time series from the training data that looks most similar to the query time series, but is in another class. To find the native guide, we need to calculate the distance between the points in the query time series and the points in native guide candidates. The candidate with the smallest distance to the query is the query's (T_q) counterfactual nearest (unlike) neighbour (T_{NUN}). It is recommended to use Dynamic Time Warping (DTW) as a distance metric [5]. Unlike the euclidean distance, that compares every two points of two sequences at the same time stamps, DTW compares the differences of two sequences by aligning an optimal warping path using sampled points in time and a cost matrix. The more similar two points (x_q, x_{NUN}) are, the lower their cost $c(x_q, x_{NUN})$ [58][59]. Figure 3.4 shows such an exemplary warping path.

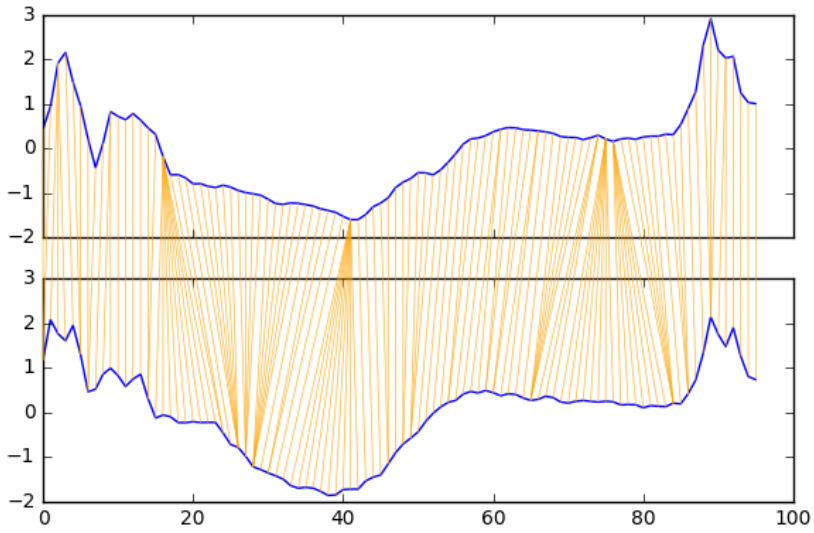


Figure 3.4.: Optimal Warping Path for two Sequences from ECG200

3.1.4. Perturbation

The last step is the generation of the counterfactual T' by performing a perturbation between the query T_q and the Native Guide T_{NUN}. That means, a subsequence S_q in T_q is swapped by a feature important subsequence S_{NUN} of T_{NUN} that is provided by the class activation mapping. In the beginning, the length l_s of the swapped subsequences is very small, e.g. $l_s = 1$, and we increase l_s, until the generated T' is classified into the desired counterfactual class $c' \in Y$. To be able to do this, all data must have the same length. We illustrated this step in algorithm 2.

Algorithm 2 Generation of $T^{'}$ by Perturbation

1: $l_s \leftarrow 1$
2: $c \leftarrow class(T_q)$
3: $c^{'} \leftarrow class(T_{NUN})$
4: **do**
5: $\quad startingPoint \leftarrow getSubsequence(T_{NUN}, l_s)$
6: $\quad T^{'} \leftarrow swap(T_q, T_{NUN}, startingPoint, l_s)$
7: $\quad l_s \leftarrow l_s + 1$
8: **while** $class(T^{'}) = c \vee \text{class}(T^{'}) \neq c^{'}$
9: **return** T'

class() returns the class of an input.
getSubsequence() returns the starting point of the most influential subsequence.
swap() swaps the datapoints of 2 inputs at a starting point and length.

3.2. Investigation and Observation of the Method

This chapter investigates the method based on diverse points of interests that could affect counterfactual goodness to gain more insights about the Native Guide's workings and also collects observations that give a foundation to propose hypotheses about potential improvements in the next chapter. First, we compare three different potential classifiers in chapter 3.2.1. Second, we investigate the influence of ECG signal strength and wavelength in 3.2.2. Next, we investigate the swapped subsequence-length in chapter 3.2.3 and influence of data quantity, length and variety in 3.2.4. Lastly, we compare the influence of different decision boundaries in 3.2.5. Because PTB-XL represents the most extensive and state-of-the-art multi-class, multi-label, and multivariate dataset, we only used that for the investigations.

3.2.1. Comparison of Classifiers

As introduced in chapter 3.1.1, we compare three different convolutional neural networks as possible classifiers for ECG classification and the Native Guide method: FCN, ResNet, and InceptionTime. Because Native Guide requires normalized data, we fine-tuned the model on the normalized PTB-XL dataset. The data was first divided into training, validation, and testing sets using a stratified k-fold with $k < 8$ for training, $k = 9$ for validation and $k = 10$ for testing data. Stratified k-fold divides the dataset into k folds. Rather than the splits being completely random, the distribution into the classes is preserved for every fold [60]. After the clean up of illogical class combination, we ended up with having 16758 normalized training and 2115 normalized validation data samples for the fine-tuning. Then we evaluated the accuracy using 2112 normalized testing data. The normalization process is explained later in chapter 3.3.3.

All three architectures have already been implemented and evaluated on the PTB-XL dataset in [47]. Their experiment on superdiagnosis classification produced different labels when using different architectures. Macro averaged accuracy reduces multiclass predictions down to multiple sets of binary predictions, calculates the corresponding metric for each of the binary cases, and then averages the results together. We use the area under the curve (AUC) as a performance measurement, which was also used by the paper [47] where we have the models from. AUC tells how much a model can distinguish between classes. Here are the evaluated AUC results for the three approaches from [47] and our normalized and fine-tuned models:

Classifier	**AUC**
FCN	0.925
ResNet	0.930
InceptionTime	0.921

Table 3.1.: Classifier Accuracy Results [47]

Classifier	**AUC**
FCN	0.9130
ResNet	0.9161
InceptionTime	0.8894

Table 3.2.: Classifier Accuracy Results (Normalized Data)

Comparing the results to the work of N. Strodthoff [47], the respective accuracies for the evaluation with normalized data are slightly lower. Nonetheless, the corresponding classifier for the maximum and minimum accuracy remained equal. ResNet seems to provide the most accurate classifications, whereas InceptionTime even dropped under 90%. Still, all classifiers have very similar performance values, so there seems to be no clear winner. Moreover, the original paper [5] used the FCN as the classifier of choice. We also used that architecture for the next investigations.

Furthermore, the generation of counterfactuals can vary when using different classifiers. One difference can be the selection of another native guide, because the classifier predicted other classes for the native guide candidates. Moreover, the length of the swapped subsequence S_{NUN} can be shorter or longer, because the predicted values for the counterfactual vary with different classifiers.

Lastly, model accuracy does not only influence the predictions, but also plausible counterfactual generation. Every step of the Native Guide method uses the classifier inside their algorithms. Thus, if the accuracy of the model is not good enough, native guide candidates or the counterfactual could be misclassified. For instance, swapping during perturbation could end too soon or too late, because the counterfactual is not classified correctly. High model accuracy is an important requirement for good counterfactual generation using the Native Guide method.

3.2.2. ECG Signal Strength and Wavelength

Different time series data has different characteristics. Some may be easy to understand, analyse, and generalize, whereas others are more complex. Even datasets from the same domain can be very diverse. The characteristics of ECG make them hard to analyse and use for algorithmic methods like Native Guide. Even though every heartbeat is schematically similar, they still have more or less different rhythms, wavelengths and signal strengths depending on aspects like age, gender, and even sleep [61].

As described in chapter 3.1.4, the native guide method swaps data points between two ECG signals. If these signals are not in sync with similar signal strength, wavelength, and concurrent peaks, perturbation results in implausible counterfactuals, because both signals do not match with each other and therefore different ECG segments are swapped. Though the method takes the nearest unlike neighbour to swap with, observation has shown that their lineament can still vary a lot. Figure 3.5 shows the first two standard leads of such implausible counterfactuals due to not synchronized data compared to the same sample with synchronized data. Differences between the query ECG and the counterfactual ECG are much clearer in sub-figure 3.5.b. A full 12-lead representation of both is provided in appendix B. Extremely high jumps between a swapped point and a neighbouring query point can also lead to implausible counterfactuals, because such anomalies do not occur in any ECG signals.

3.2.3. Swapped Subsequence-Length

During Perturbation in 3.1.4, the algorithm is taking the most influential continuous subsequence and replacing it with the subsequence at the same position from the native guide time series. Especially with very long data, counterfactuals also tend to be very long. It does not only swap specific important points from the series, but whole sequences that may include important and non-important points. Particularly for longer ECG data with repetitive behavior and anomalies occurring once every cycle, the whole query time series may have to be swapped to remove a characteristic disease.

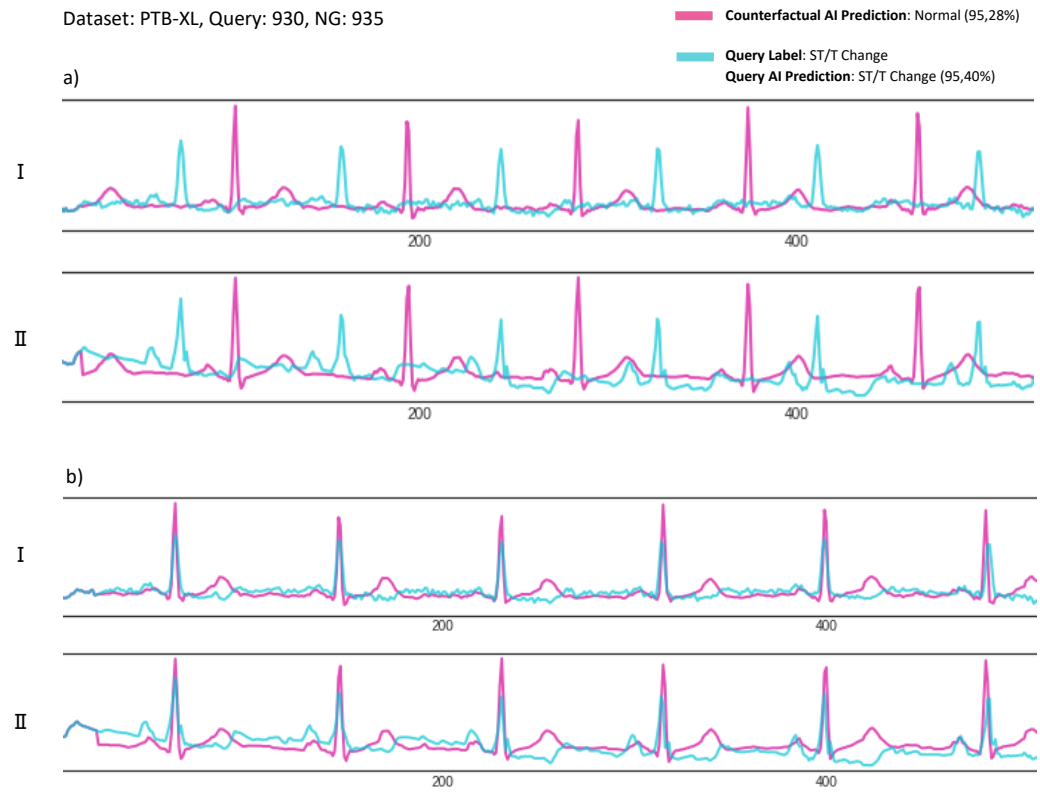


Figure 3.5.: a) Implausible Counterfactual, b) Plausible Counterfactual

3.2.4. Data Quantity, Length and Variety

The available amount of data has strong influence on the algorithms of Native Guide. The proposed method uses the training data as native guide candidates. The more training data, the more candidates and the longer it takes to find the nearest native guide, as described in 3.1.3. Conversely, only few native guide candidates provide less coverage for all discriminative areas of an inspected domain, moreover, causing problems with model training [5][50]. The goal is to find a balance between the amount of native guide candidates that covers enough data space but also does not increase the processing time. To produce counterfactuals for the PTB-XL dataset, we reduced the number of candidates and smoothed class imbalance by randomly selecting 200 ECG samples from each class, which results in remaining 1000 data native guide candidates for the counterfactual generation process.

Likewise, data length appears to be a key factor for Native Guide's implementation requirements and performance. The longer the time series the longer it takes to perform the perturbation step from section 3.1.4. Besides that, shorter time series could not show recurrent characteristics like periodicity, seasonality, rhythm, and, particularly in case of ECG data, heart rate. If the

classifier does not learn these characteristics, because it has not seen them in the data, the perturbation step could produce counterfactuals violating these characteristics and the classifier would not be able to detect the problem.

Lastly, multivariate data can be challenging for Native Guide. Usually, class activation mapping does not consider multiple dimensions in the time series. That means, the importance values represent all data points across all dimensions at a given time. Multi-lead ECGs measure heart impulses from different body regions at the same time. Therefore, the leads in one ECG are very similar with an aligned waveform and equal heart rate. Swapping the same point in every lead with synchronized points from the counterfactual time series makes rather sense. Multivariate time series from other domains may be more chaotic with more autonomous leads. The practicality of Native Guide with such data would require further research.

3.2.5. Different Decision Boundaries

The original approach [5] used 50 % as the decision boundary, which means that a counterfactual gets approved when the model is with a prediction over 50 % confident that the counterfactual is not a member of its previous class. With just one percent more the counterfactual would still be assigned to that class. Because the area around this 50 % is very ambiguous, it is interesting how a shifted decision boundary changes characteristics like confidence in the previous class, confidence in the new class, and also the length of the subarray S_{NUN}, that needs to be swapped with the native guide to assign the counterfactual to an opposite class. Therefore, we shifted the decision boundary between 50 % and 10 % for 14 randomly chosen query time series from the PTB-XL dataset and collected each confidence score and changed subarray-length. Taking the means for each boundary, we can draw figure 3.6. The mean confidence scores describe, how sure the model is that the counterfactual ECG is in a specific class. We distinguish between the confidence of the previously assigned class, which is directly influenced by the decision boundary value and the newly assigned class.

It is inevitable that the confidence score of the previous class rises when the decision boundary is higher. In that case, the lower the confidence score, the less risky is the counterfactual, because the model is more confident that the counterfactual is not a member of this class anymore. On the other hand, the confidence score of the newly assigned class or classes rises imperceptibly the lower the decision boundary. Conversely, the higher the confidence score, the more certain is the counterfactual, because the model is more confident of the new class membership. However, it is important to mention, that at first with some instances is was not possible to generate a counterfactual with the decision boundary at 10 % or 20 %, because baseline native guide's confidence of the query's class is already higher than that. As a result, even if all data

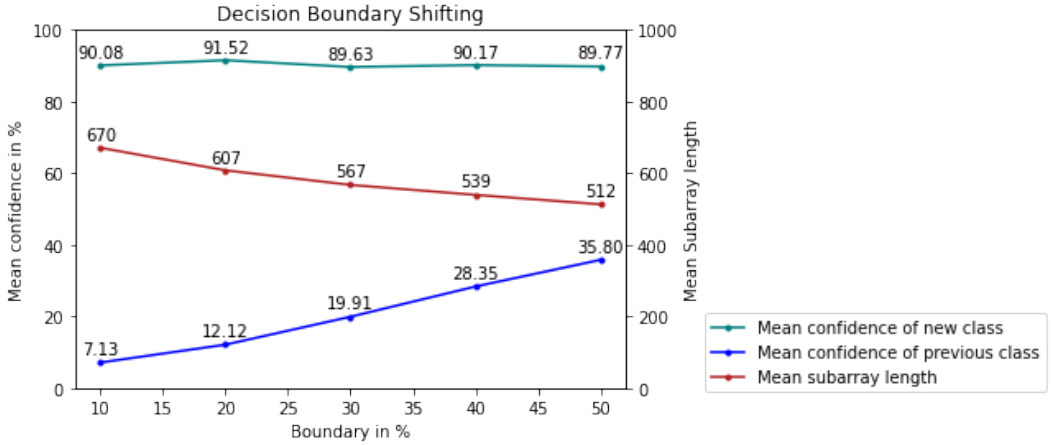


Figure 3.6.: Decision Boundary Shifting

points are swapped, the decision boundary still is not met. Because of that, we only select native guides with respect to a decision boundary at 10 %, which is the minimal boundary for our experiment.

3.3. Experimental Approaches for Optimization

In the following section we will see three different experiments that introduce optimization approaches to handle the downsides of the Native Guide method, but still keep the basic idea of the original approach, which is swapping data points with a nearest unlike neighbour [62]. First, we propose a normalization and synchronization of ECG data in section 3.3.1. Second, we test another approach for perturbation swapping points instead of subsequences in 3.3.2. Third, we introduce a shifted desicion boundary in chapter 3.3.3.

3.3.1. Normalization and Synchronization

Hypothesis

To provide more meaningful counterfactuals, ECG signals need to be normalized in signal strength and synchronized in wavelength and peak concurrency.

Proof of Concept

Algorithm 3 summarizes all the following ideas. First, we normalize the signal strength of the ECG data so that the minimum value is at zero and the maximum value at one. Algorithm 4 is provided in the appendix C that implements the normalization. Differences in signal strength between the leads of multivariate time series are kept, because normalization should be performed on all leads and not on every lead individually. The shape of the time series

Algorithm 3 Normalization and Synchronization

1: $T_q \leftarrow norm(T_q)$
2: $T_{NUN} \leftarrow norm(T_{NUN})$
3: $T_q, T_{NUN} \leftarrow waveSync(T_q, T_{NUN})$
4: $T_q, T_{NUN} \leftarrow shift(T_q, T_{NUN})$
5: **return** T_q, T_{NUN}

remains the same, however, changed values can also change the model predictions. Therefore, fine-tuning of the classification model using the normalized data is required to maintain high prediction accuracy.

The next step is the wavelength synchronization, which is more challenging and a research area itself [63][64]. To synchronize a query time series and its native guide time series, we first need to find the wavelength of both. We define the wavelength as the length between two heartbeat R-peaks. Identifying the right peaks in an ECG sample can be difficult, since signal strength can be chaotic and vary [61]. After calculating the wavelengths, the native guide time series needs to be scaled or squished to match the wavelength of the query time series. Algorithm 5 in appendix C implements that idea.

Lastly, a time series is shifted, so that the peaks of both are concurrent. This can be achieved by calculating the fast-cross-correlation using Fourier-transformations [65]. It recognizes the frequencies of periodic signals by spinning them around a circle and averaging the sampled points. This average is maximal, when the signal is wrapped around with the right frequency [65][66]. The algorithm 6 in appendix C shows a more detailed exemplary solution [65][67].

It is important to mention that the wavelength synchronization and shifting alter the length of the time series. They get aligned by shortening the longer time series. Furthermore, wavelength synchronization only changes the wavelength of the native guide sample and not the query time series. Figure 3.7 illustrates, how a synchronization between a query time series (blue) and a native guide (pink) looks like.
A significant question is whether the synchronization could alter the Native Guide ECG in a way that it gets assigned to another class. In that case, only native guides without model prediction change can be considered. However, an optimal synchronization should only change the wavelength of a time series. According to the study of G. Quer et al. [61], the heart rate of two healthy individuals can differ by up to 70 beats per minute. Despite this, healthy hearts are considered to have between 50 and 100 beats per minute [61][68].

More research is needed to evaluate, whether synchronization of two ECG time

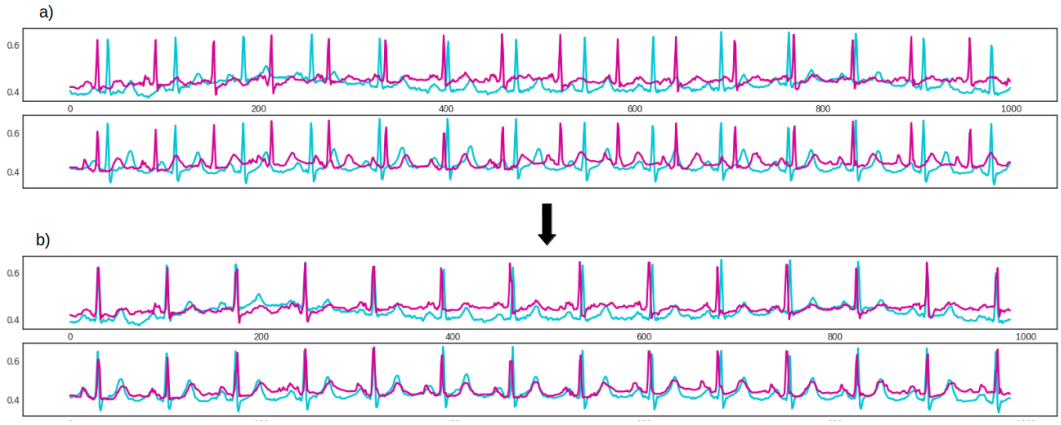


Figure 3.7.: ECG Synchronization: a) Raw, b) Synchronized

series can have a significant impact on the time series class membership. As a first test, we compared 2112 ECG time series predictions before and after synchronization. The data in the test is from the PTB-XL dataset and the classifier is a FCN with multi-label predictions [47]. 2112 ECGs from the test dataset split where synchronized with their closest native guide out of 1000 randomly chosen native guide candidates from the training dataset split with equally distributed 200 samples for each class. For all 2112 native guides, the synchronization changed the model's prediction of in total 129 synchronized native guides, resulting in a 93.86% success rate in not changing the predicted label after synchronization.

Conclusion

All things considered, a risk-free normalization and synchronization strongly depends on a high-performing algorithms and a good classifier. That classifier should ignore unimportant changes but be able to detect real issues arising from heart rate changes produced by a synchronization. This way it could prevent the algorithm from using such native guides. However, counterfactuals of normalized and synchronized data usually look more plausible than counterfactuals of not normalized or synchronized data, which means that conspicuous segments that cause the prediction can be easier identified with the synchronization.

3.3.2. Swapping Points instead of Subsequences

Hypothesis

Swapping important points instead of subsequences reduces the total amount of swapped data points and still produces plausible counterfactuals.

Proof of Concept

We changed the perturbation algorithm from chapter 3.1.4 so that it takes the indices of the feature map weights in descending order by value. Then, data points are swapped one after another, starting with the corresponding highest weight. That way, the most influential points are swapped first until the generated counterfactual T' changes its prediction to the native guide's class. Algorithm 7 in appendix C illustrates the procedure.

Producing ten counterfactuals for ten random data samples, we observed time series that go back and forth between the query time series and its native guide, resulting in very implausible chaotic counterfactuals, that do not reflect any real ECG signals. Figure 3.8.a shows a part of such a counterfactual in comparison to changing the whole subsequence in 3.8.b. Apart from that, we compared the amount of swapped data points between point and subsequence swapping within 10 randomly selected samples with a decision boundary at 10 %. On average swapping points changed 863 and swapping one subsequence changed 716 data points.

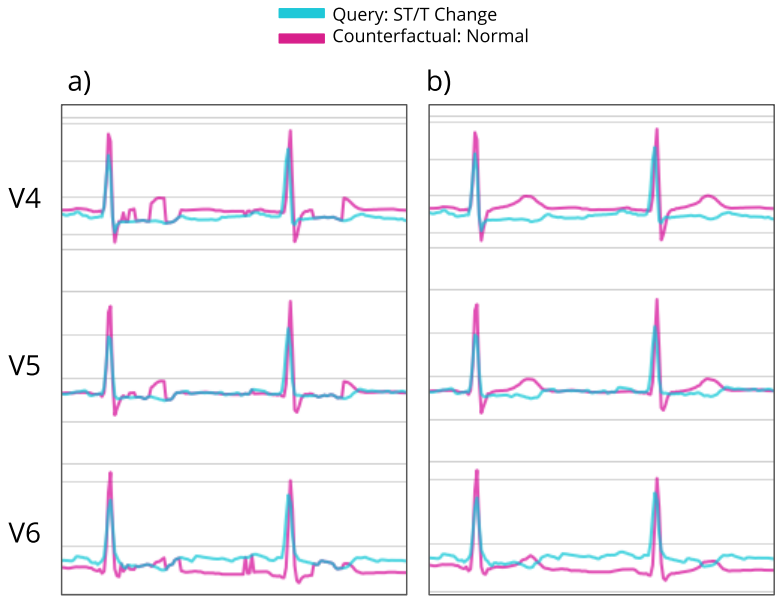


Figure 3.8.: Perturbation Approach: a) Points, b) Subsequences

Conclusion

Considering the implausibility of counterfactuals produced by point for point perturbation and, as it turns out, even higher amount of swapped data points, this approach does not provide any improvement of the original perturbation

method of swapping the most important subsequence. A theory that explains the higher amount of swapped points is that the implausibility of the counterfactuals causes problems of classifying it. On the other hand, subsequence counterfactuals actually almost represent the corresponding synchronized native guides, because so many points have to be swapped to change the prediction to the desired outcome.

3.3.3. Shifted Decision Boundary

Hypothesis

A decision boundary lower than 50 % reduces the risk of generating a wrongly classified counterfactual, which would be a sample that is classified to the target class by a classifier but in reality does not belong to that class.

Proof of Concept

Chapter 3.2.5 has shown how some samples react to shifted decision boundaries that are lower than 50 %. We use the prediction probabilities as confidence scores of the model. For decision boundaries at 50 % to 10 %, the mean confidence score of the newly predicted class decreases very slightly from 90.08 % to 89.77 %, while the confidence of the previous class rapidly decreases from 7.13 % to 35.80 %. This supports the stated hypothesis. It can also be seen that together with the decline of the decision boundary, the mean subarray length S_{NUN} rises from 512 to 670.

Conclusion

The conducted test shows that the risk of generating wrongly classified counterfactuals indeed can be reduced by lowering the decision boundary threshold, because the probability of previous classes decrease while probability of the new classes increase. However, this also means that the counterfactual is less close to the query time series and therefore shrinks the proximity, which describes how close the counterfactual is to its query time series. The problem of not reaching the decision boundary can be solved by previously selecting native guides that hold more suitable confidence values. To find a trade-off between less risky but more proximate counterfactuals, a decision boundary at 40 % seems like a good solution. That way ambiguous predictions around 50 % are obviated.

4. Evaluation: Expert Interview

For evaluation and knowledge-seeking purposes, we conducted an expert interview with cardiologists. The following chapters describe its goal, structure, and approach in chapter 4.1, expert background in 4.2 and finally interview results in section 4.3.

4.1. Goal, Structure and Approach

Expert interviews are suitable research methods when dealing with complex problems with the necessary context knowledge provided by any experts. It summarizes and confirms research results and helps to learn more about possible solutions to the problem [69][70].
ECG classification is a very complex problem and only experts can reliably evaluate data and usefulness of a counterfactual method in that area. The problem-centered expert interviews achieve the following goals: First, obtaining more knowledge about the practical use of ECGs in the medical field. Second, gaining insights about the plausibility of counterfactuals provided by the Native Guide method. Third, exploring possible use-cases of ECG data classification with counterfactuals. Fourth, receiving improvement ideas, and last, finding out how experts think about the idea of counterfactuals in general.
The interviews took between 45 and 60 minutes, were semi-structured and involved mostly open questions. Such an approach guides the interviewee thematically, but also allows more freedom to answer and ask spontaneous questions. The audio of the interviews were recorded. Then it was transcribed and analysed using a structured qualitative analysis approach introduced by Mayring [71].

After explaining the background and goals of the interview, introductory questions were asked to get to know the expert's experience, intentions and needs and therefore already lay a foundation for the next steps. The next part was conceived as a usability testing with the specific goal to evaluate the design, usefulness, plausibility, and satisfaction of the counterfactual explanation method on ECG data classification from a real user's perspective with exemplary data. Moreover, it provided expert improvement ideas that can help to advance such methods in future works. We tested six ECG samples from the PTB-XL dataset covering at least one sample per class. We chose random samples, but required a seemly good synchronization between the queries and their native guides. To see similarities in the decisions but also to evaluate more counterfactual samples, two samples were evaluated by both experts, whereas

each expert was provided with two more samples different from the other expert. For all samples we followed the same procedure.

At first, the expert was asked to assign the sample to one of five classes and point out indicative areas. Then, the AI prediction was revealed and discussed. Next, we showed the counterfactual and discussed its plausibility. As a last step, the expert was asked to paint in an optimal counterfactual themselves. To conclude, the experts were interviewed to discover the possible use-cases for ECG classification with counterfactuals, the overall attitude towards the method, and potential improvement areas.

4.2. Expert Background

In total, two experienced cardiologists from the same clinic in Germany where consulted. One is the head doctor of different fields including internal medicine and will be called expert A. The other is a senior doctor for the echocardiography field and will be referred to as expert B. Both interviews followed the same script, but depending on the answers and motivation follow-up questions did vary.

4.3. Interview Results

All following statements can be retraced in the appendix A, where the converged structured qualitative analysis of the interviews is provided. The results are categorized into the usage of ECG, ECG data quality, general attitude towards counterfactuals, plausibility of counterfactuals, improvement ideas, and possible use-cases for the method.

4.3.1. Usage of ECG

The experts pointed out, that electrocardiography is a very important cardiological functional tool. Every patient, that is assigned to the internal medicine department because of any symptoms or specific criteria, certainly gets an ECG. It is a fundamental diagnostic method to recognize heart problems. An equipped clinic saves all ECG records to an internal computer program, were they could be easily accessed also later in time and potentially shared for projects like this thesis, when personal information of patients are withhold. The potential use of ECGs for identification or authentication can limit open source use, because ECGs qualify as personal data, which needs to be protected.

Experienced doctors classify an ECG as being normal or pathological in a few seconds. If there is a more complex or ambiguous problem, a group of experts may need up to five minutes. Both experts would restrain themselves from showing the ECG to their patients, because in their opinion it is too complex

to understand. Information overload of patients often does more bad than good.

In Germany the 12-lead ECG is the state-of-the-art way to record and analyse an ECG. Some doctors would actually prefer to have the option to see more ECG leads, if they deem it to be necessary. Nevertheless, if one would need to only use one lead, it should be lead II of the standard leads, which according to expert A, would be close to clairvoyance. It is important to mention that doctors also require a writing speed of 50 mm per second and a calibration of 1 mV per 1 cm for the amplitude of the signal strength. They are used to that specific form of display. Similarly, American doctors are used to a writing speed of 25 mm per second and have problems analysing the German standard. Lastly, expert A wishes more doctors to be capable of reliably evaluating an ECG, in case they are working close to the internal medicine field. Especially young physicians feel the lack of knowledge or experience in ECG diagnostics.

4.3.2. ECG Data Quality

According to the expert statements during the usability testing, some ECG examples from the PTB-XL dataset seem to be incorrectly classified by the AI, but also mislabeled by the doctors that assigned the classes in the first place. The false myocardial infarctions were very concerning. Misclassification of those is also a very common mistake by junior doctors in that field. As a result, AI predictions cannot be accurate in real life if the given training data is already assigned to wrong classes. This indicates the need to either revisit the dataset and correct wrong labels or provide better datasets with equal or a greater number of samples and labels assigned by more experienced cardiologists.

4.3.3. General Attitude towards Counterfactuals

The interviews with real experts revealed that counterfactuals are a very promising explanatory method for ECG classification. They have never seen such an approach before and were interested and excited for its future application, which could boost further research in that area. For a counterfactual to make sense expert A would need minimal or very subtle changes towards a threshold to be convinced of its usefulness. Referring to the views of expert B, counterfactuals need to be able to recognize the distinct characteristics of the different pathological classes. These characteristics need to be modified gradually until they are in a normal state, while the level of difficulty can vary a lot for different diseases. They need to be considered independently by providing the suitable changes for certain abnormalities.

4.3.4. Plausibility of Counterfactuals

Expert A was disappointed by some misclassified samples and therefore did not see the usefulness for some of the counterfactuals throughout the test. However, he assigned all but one counterfactual to their predicted class. Also, all but one counterfactual changed an ECG in a way, that it became normal. Only one tried to go the other way and make a pathological ECG out of a normal ECG. According to expert A, that last counterfactual actually turned out to be even more normal than its query ECG.

Expert B was constantly very pleased with the given counterfactuals and evaluated all as plausible and useful. He especially complimented the synchronized overlay of the query ECG and the counterfactual, without being prompted about it. Expert B declared that counterfactuals would emphasize the underlying causes for a specific classification quite nicely. In case of conduction disturbance or arrhythmia, where total synchronization should be impossible, it helped expert B to see the first R peaks concurrent and recognize pathological delays because of later asynchrony with the counterfactual.

Apart from that, both experts agreed that the counterfactuals look like real ECG time series. Appendix A shows all tested ECG samples with the model prediction, originally assigned label and prediction of the experts. The experts also highlighted conspicuous segments. For every ECG, the Native Guide counterfactual and counterfactuals drawn by the experts are also provided in the same appendix.

4.3.5. Improvement Ideas

Expert A stated that counterfactuals for ECG data should only show subtle indispensable changes in the most relevant segments. The greatest interest seems to be the critical threshold where ECGs are either physiological or pathological. Therefore a counterfactual should represent exactly that threshold. This would still need to be incorporated in the Native Guide method for ECG classification. Synchronization is already a beneficial step towards that goal. Still, the changed subsequence remains very long and crude. However, expert B did not mention the need for any more subtle changes.

4.3.6. Possible Use-Cases

From consultation with the experts there are various use-cases for ECG counterfactuals. First, they can be useful for the education and training of students and junior doctors. Generally, beginners learn to distinguish between different abnormalities and normal ECGs by recognizing critical sections with all possible variations and assess them correctly. That is exactly what counterfactuals are doing. Thus, counterfactuals could help students during their training phase,

which should be tested in future works with actual ongoing candidates for the internal medicine field.

Besides that, emergency doctors, like orthopedist, that are not sufficiently trained in ECG evaluation, could need reliable AI classification with counterfactuals, because it is crucial to recognize acute problems like heart attacks that need to be treated immediately. Next, assistance with AI classification is helpful in all ECG evaluation tasks where a doctor depends on measurement values, like the length of a RQ-wave. Lastly, there are several heart diseases that are difficult for many doctors to recognize or distinguish. That includes complex rhythm disturbances and storage disorders like amyloidose or morbus fabry. Changes in the ST-segment are also hard to diagnose and are often wrongly classified by the automatic analysis of the systems that are used nowadays. Expert B mentioned that he welcomes the option to show a counterfactual to any doctor that has a need or interest to see it.

5. Discussion

Even though the opportunities with Native Guide appear promising there are potential drawback with this method. Given that the technique involves perturbation of a query time series until it enters another class, the counterfactual by design is usually very close to the decision boundary. This was described as a desired property, by E. Delaney [5] as well as expert A, because the counterfactual would also be as close to the query as possible. This means that the generated data likely remains in the distribution and reveals the most important characteristics that would have to be changed to acquire the counterfactual outcome. However, a time series that is very close to a decision boundary often does not have a considerably high prediction confidence. The decision boundary is the region of a problem space in which the output class of a classifier is ambiguous. We point out that designing the generation process so close to the boundary, also means risking many misclassified counterfactuals, particularly when using inaccurate classifiers. Experiments in [72] have shown that wrongly classified data are on average significantly closer to the decision boundary than correctly classified data. There needs to be a balance between proximate counterfactuals and low risk classification. Slight decision boundary shifting, as described in chapter 3.3.3, could be one possible solution for the generation of less risky counterfactuals. The in 3.3.2 introduced idea to only swap the most important points does not provide plausible counterfactuals yet, but is a step towards the direction of producing more proximate counterfactuals.

Another aspect, that became clear during both interviews, is the appropriate plotting of the ECGs. Different countries use different calibrations of writing speed and amplitude. This leads to different ECG appearances and difficulties evaluating unfamiliar calibration styles. The specific calibration style used by the target group needs to be considered before any further testing or applications are done.

The general complexity of different time series data makes it difficult to produce good counterfactual explanations. The reference method as it is proposed by E. Delaney et al. [5] did not produce good counterfactuals for the complex ECG dataset PTB-XL. However, normalizing and synchronizing the query ECG and its native guide is one key idea that enables the generation of plausible counterfactuals in this thesis. A good synchronization algorithm should reliably spot all R-peaks and determine the average wavelength between two of such peaks. Every ECG has unique features [73]. Thus, records vary a lot and peaks can sometimes be very shallow. It is challenging to provide an algo-

rithm that ignores irrelevant noise and small peaks, like the T-wave, while at the same time finds all peaks in ECGs with low signal strength. The algorithm should make these detections whether the patient is healthy or sick.

In conclusion, the expert interviews provided key insights about understanding ECG data and analysis. They have shown that the at first most promising, comprehensive, and suitable dataset PTB-XL could have many mislabeled samples, mostly for the myocardial infarction class. This is unfortunate, because an intelligent system cannot work accurately, if underlying labels, which are required by an AI to connect them with specific features, are wrong. There is a need for better labeled ECG datasets openly accessible to enhance development and research in that area. The fact that ECGs could be used in biometric identification mechanisms makes enhancing data availability more difficult [73]. Data augmentation is a possible solution for that problem, because it generates synthetic ECG data. In fact, the Native Guide method generates ECG data making it a viable option for data augmentation [5]. However, the method would need to be altered to change enough from the original signal in order to make identification impossible.

6. Conclusion and Future Work

Despite ECG's highly complex data characteristics, it is possible to generate plausible counterfactuals for model predictions using the Native Guide method, even though the proximity factor is still capable of improvement.

To implement this method, background knowledge about machine learning, convolutional neural networks, and explainability is required. Depending on the field of application, knowledge about different data characteristics may also be needed. In a field like ECG that involves periodic time series data, synchronization has significant influence on the plausibility of counterfactuals. However, it still needs to be evaluated whether synchronization could critically alter a time series and if it does how could this problem be resolved. An example is provided that implements the synchronization step using the idea of Fourier- transformations and cross-correlation. Furthermore, perturbation with point for point swapping does not provide plausible counterfactuals, but swapping sequences on multiple repeated segments in the series could be an alternative approach. This should be implemented and tested in future work. Additionally, another interesting topic for future research would be the improvement of the diversity criteria by suggesting different counterfactuals using the next closest native guides.

To summarize, the experts have shown a willingness to use sophisticated counterfactual methods in their cardiologic field for different use cases like education. Trust in the model's predictions was increased by the use of counterfactuals, especially for expert B. All things considered, AI classification with counterfactuals has significant potential for the future of ECG analysis by improving the identification and understanding of life threatening heart diseases.

Bibliography

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug 2021. [Online]. Available: https://doi.org/10.1038/s41586-021-03819-2

[2] E. Gibney, "Google ai algorithm masters ancient game of go," *Nature*, vol. 529, no. 7587, pp. 445–446, Jan 2016. [Online]. Available: https://doi.org/10.1038/529445a

[3] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, *Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges*, 09 2019, pp. 563–574.

[4] A. Soofi and A. Awan, "Classification techniques in machine learning: Applications and issues," *Journal of Basic Applied Sciences*, vol. 13, pp. 459–465, 08 2017.

[5] E. Delaney, D. Greene, and M. T. Keane, "Instance-based counterfactual explanations for time series classification," *CoRR*, vol. abs/2009.13211, 2020. [Online]. Available: https://arxiv.org/abs/2009.13211

[6] N. Roese and M. Morrison, "The psychology of counterfactual thinking," *Historical Social Research*, vol. 34, 01 2009.

[7] R. Byrne, "Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning," 08 2019, pp. 6276–6282.

[8] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," 2018.

[9] M. T. Keane, E. M. Kenny, E. Delaney, and B. Smyth, "If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques," 2021.

[10] E. Ates, B. Aksar, V. J. Leung, and A. K. Coskun, "Counterfactual explanations for machine learning on multivariate time series data," 2020.

[11] I. Karlsson, J. Rebane, P. Papapetrou, and A. Gionis, "Explainable time series tweaking via irreversible and reversible temporal transformations," ser. 2018 IEEE International Conference on Data Mining (ICDM); International Conference on Data Mining Proceedings, 2018, pp. 207–216. [Online]. Available: http://urn.fi/URN:NBN:fi:aalto-201901141165

[12] S. Tonekaboni, S. Joshi, D. Duvenaud, and A. Goldenberg, "Explaining time series by counterfactuals," 2020. [Online]. Available: https://openreview.net/forum?id=HygDF1rYDB

[13] A. H. Gee, D. Garcia-Olano, J. Ghosh, and D. Paydarfar, "Explaining deep classification of time-series data with learned prototypes," 2019.

[14] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, *'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions.* New York, NY, USA: Association for Computing Machinery, 2018, p. 1–14. [Online]. Available: https://doi.org/10.1145/3173574.3173951

[15] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 275–285. [Online]. Available: https://doi.org/10.1145/3301275.3302310

[16] C. Fernández-Loría, F. Provost, and X. Han, "Explaining data-driven decisions made by ai systems: The counterfactual approach," 2021.

[17] R. Wieringa, *Design science methodology for information systems and software engineering.* Netherlands: Springer, 2014, 10.1007/978-3-662-43839-8.

[18] M. Robeer, "Contrastive explanation for machine learning," Master's thesis, Utrecht University, 7 2018.

[19] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," 2016.

[20] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, p. 1936–1962, Sep 2020. [Online]. Available: http://dx.doi.org/10.1007/s10618-020-00710-y

[21] J. Han, M. Kamber, and J. Pei, "13 - data mining trends and research frontiers," in *Data Mining (Third Edition)*, third edition ed., ser. The Morgan Kaufmann Series in Data Management Systems, J. Han,

M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 585–631. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123814791000137

[22] J. C. B. Gamboa, "Deep learning for time-series analysis," *CoRR*, vol. abs/1701.01887, 2017. [Online]. Available: http://arxiv.org/abs/1701.01887

[23] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Mining and Knowledge Discovery*, vol. 33, no. 4, p. 917–963, Mar 2019. [Online]. Available: http://dx.doi.org/10.1007/s10618-019-00619-1

[24] P. Schaefer and U. Leser, "Multivariate time series classification with weasel+muse," 2018.

[25] X. Wang, K. Smith-Miles, and R. Hyndman, "Characteristic-based clustering for time series data," *Data Min. Knowl. Discov.*, vol. 13, pp. 335–364, 09 2006.

[26] A. Berg, T. McMurry, and D. N. Politis, "2 - testing time series linearity: Traditional and bootstrap methods," in *Time Series Analysis: Methods and Applications*, ser. Handbook of Statistics, T. Subba Rao, S. Subba Rao, and C. Rao, Eds. Elsevier, 2012, vol. 30, pp. 27–42. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444538581000028

[27] O. Rose, "Estimation of the hurst parameter of long-range dependent time series," 1996.

[28] Z. Liu, "Chaotic time series analysis," *Mathematical Problems in Engineering*, vol. 2010, 03 2010.

[29] D. Poole and A. Mackworth, *Artificial Intelligence: Foundations of Computational Agents*, 2nd ed. Cambridge, UK: Cambridge University Press, 2017. [Online]. Available: http://artint.info/2e/html/ArtInt2e.html

[30] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. The MIT Press, 2012.

[31] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[32] M. Kumar and S. Dargan, "A survey of deep learning and its applications: A new paradigm to machine learning," *Archives of Computational Methods in Engineering*, 06 2019.

[33] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. V. Esesn, A. A. S. Awwal, and V. K. Asari, "The history began from alexnet: A comprehensive survey on deep learning approaches," 2018.

[34] Z. Meng, Y. Hu, and C. Ancey, "Using a data driven approach to predict waves generated by gravity driven mass flows," *Water*, vol. 12, 02 2020.

[35] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural Networks*, vol. 1, pp. 119–130, 1988.

[36] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, 01 2012.

[37] M. D. Pra. (2020, 09) Time series classification with deep learning. [Online]. Available: https://towardsdatascience.com/time-series-classification-with-deep-learning-d238f0147d6f

[38] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/18

[39] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 0210–0215.

[40] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, Jun. 2019. [Online]. Available: https://ojs.aaai.org/index.php/aimagazine/article/view/2850

[41] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519308103

[42] A. C. Quelhas, C. Rasga, and P. N. Johnson-Laird, "The relation between factual and counterfactual conditionals," *Cognitive Science*, vol. 42, no. 7, pp. 2205–2228, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12663

[43] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association

for Computing Machinery, 2020, p. 607–617. [Online]. Available: https://doi.org/10.1145/3351095.3372850

[44] S. Jambukia, V. Dabhi, and H. Prajapati, "Classification of ecg signals using machine learning techniques: A survey," 03 2015.

[45] R. Sinha, "An approach for classifying ecg arrhythmia based on features extracted from emd and wavelet packet domains," Ph.D. dissertation, 07 2012.

[46] "cardiosecur magazine ecg lead systems," https://www.cardiosecur.com/magazine/specialist-articles-on-the-heart/lead-systems-how-an-ecg-works, accessed: 2021-08-11.

[47] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ecg analysis: Benchmarks and insights from ptb-xl," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, 2021.

[48] K. Buza, A. Nanopoulos, L. Schmidt-Thieme, and J. Koller, "Fast classification of electrocardiograph signals via instance selection," in *2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology*, 2011, pp. 9–16.

[49] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh, "The ucr time series archive," 2019.

[50] V. Nasir and F. Sassani, "A review on deep learning in machining and tool monitoring: methods, opportunities, and challenges," *The International Journal of Advanced Manufacturing Technology*, vol. 115, 08 2021.

[51] P. Matias, D. Folgado, H. Gamboa, and A. Carreiro, "Robust anomaly detection in time series through variational autoencoders and a local similarity score," 01 2021, pp. 91–102.

[52] S. Fazeli. (2018, 05) Ecg heartbeat categorization dataset. [Online]. Available: https://www.kaggle.com/shayanfazeli/heartbeat

[53] A. Goldberger, L. Amaral, L. Glass, S. Havlin, J. Hausdorg, P. Ivanov, R. Mark, J. Mietus, G. Moody, C.-K. Peng, H. Stanley, and P. Physiobank, "Components of a new research resource for complex physiologic signals," *PhysioNet*, vol. 101, 01 2000.

[54] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," 2014.

[55] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[57] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014.

[58] G. Forestier, F. Petitjean, H. Dau, G. Webb, and E. Keogh, "Generating synthetic time series to augment sparse datasets," in *Proceedings*, V. Raghavan, S. Aluru, G. Karypis, L. Miele, and X. Wu, Eds. United States of America: IEEE, Institute of Electrical and Electronics Engineers, Dec. 2017, pp. 865–870, iEEE International Conference on Data Mining 2017, ICDM 2017 ; Conference date: 18-11-2017 Through 21-11-2017. [Online]. Available: http://icdm2017.bigke.org/ ,https://ieeexplore.ieee.org/xpl/conhome/8211002/proceeding

[59] *Dynamic Time Warping*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 69–84. [Online]. Available: https://doi.org/10.1007/978-3-540-74048-3_4

[60] F. Farias, T. Ludermir, and C. Bastos-Filho, "Similarity based stratified splitting: an approach to train better classifiers," 2020.

[61] G. Quer, P. Gouda, M. Galarnyk, E. J. Topol, and S. R. Steinhubl, "Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, bmi, and time of year: Retrospective, longitudinal cohort study of 92,457 adults," *PLOS ONE*, vol. 15, no. 2, pp. 1–12, 02 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0227709

[62] S. L. Pfleeger, "Experimental design and analysis in software engineering," *Annals of Software Engineering*, vol. 1, no. 1, pp. 219–253, Dec 1995. [Online]. Available: https://doi.org/10.1007/BF02249052

[63] B. Du, R. Deng, and X. Sun, "Precise frequency synchronization detection method based on the group quantization stepping law," *PLOS ONE*, vol. 14, no. 2, pp. 1–12, 02 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0211478

[64] M. Khorasani. (2021) How to synchronize time series datasets in python. [Online]. Available: https://towardsdatascience.com/how-to-synchronize-time-series-datasets-in-python-f2ae51bee212

[65] L. Fridman, D. E. Brown, W. Angell, I. Abdić, B. Reimer, and H. Y. Noh, "Automated synchronization of driving data using vibration and steering events," *Pattern Recognition Letters*, vol. 75, p. 9–15, May 2016. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2016.02.011

[66] S. Riffle. (2012) Understanding the fourier transform. [Online]. Available: https://www.i-programmer.info/programming/theory/3758-understanding-the-fourier-transform.html

[67] L. Fridmann. (2015) Fast cross correlation and time series synchronization in python. [Online]. Available: https://lexfridman.com/fast-cross-correlation-and-time-series-synchronization-in-python/

[68] R. H. Shmerling. (2020, 03) How's your heart rate and why it matters? [Online]. Available: https://www.health.harvard.edu/heart-health/hows-your-heart-rate-and-why-it-matters

[69] S. Döringer, "The problem-centred expert interview. combining qualitative interviewing approaches for investigating implicit expert knowledge," *International Journal of Social Research Methodology*, vol. 24, no. 3, pp. 265–278, 2021. [Online]. Available: https://doi.org/10.1080/13645579.2020.1766777

[70] S. Lichtenstein. (2019) Ux in the wild: Expert interview. [Online]. Available: https://medium.com/ux-in-the-wild/ux-in-the-wild-expert-interview-b571d7137964

[71] P. Mayring, "Qualitative inhaltsanalyse," in *Handbuch qualitative Forschung : Grundlagen, Konzepte, Methoden und Anwendungen*, U. Flick, E. v. Kardoff, H. Keupp, L. v. Rosenstiel, and S. Wolff, Eds. München: Beltz - Psychologie Verl. Union, 1991, pp. 209–213.

[72] D. Mickisch, F. Assion, F. Greßner, W. Günther, and M. Motta, "Understanding the decision boundary of deep neural networks: An empirical study," *CoRR*, vol. abs/2002.01810, 2020. [Online]. Available: https://arxiv.org/abs/2002.01810

[73] S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, and B. K. Wiederhold, "Ecg to identify individuals," *Pattern Recognition*, vol. 38, no. 1, pp. 133–142, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320304002419

Appendix

A. Expert Interview Analysis and ECG Plots

Pre-Task Interview		
Textabschnitt	**Kodierung**	**Kategorie**
Experte A: "Also in der Inneren Medizin sollte es so sein, dass jeder Patient, der sich notfallmäßig selbst vorstellt, in einer Notaufnahme im Krankenhaus oder vorgestellt wird, ein EKG bekommt. (...) Also jeder, jeder, jeder Patient, der der Inneren zugeordnet wird. (...) wenn man zusätzlich zu dem ersten EKG ein weiteres EKG schreibt, das auch immer dann, wenn Beschwerden auftreten, die hinweisend für eine Herzerkrankung sein könnten." ***Experte B:*** "Also wir setzen das EKG ständig ein, weil das meines Wissens nach der wichtigste Baustein ist im Rahmen der nicht invasiven kardiologischen Funktionsdiagnostik. Da baut sich alles andere als erst auf. Bei uns wird also denke ich mal oder sollte bei jedem Patienten ein EKG zur Aufnahme geschrieben werden, wenn er hier über die Notaufnahme kommt oder auch direkt in die Klinik aufgenommen wird. In der Ambulanz ohnehin."	Anwendung von EKGs im Krankenhaus	Anwend-ungsfall
Experte A: "Wir arbeiten hier mit der EDV-Übertragung, also praktisch mit Schiller System und mit Cema 200 am PC. Können uns dann das EKG hier darstellen, (...) digital schon hier, auf dem auf dem Computer."	Genutzte Technik zur EKG Aufnahme und Darstellung	Vorgehen mit EKGs
Experte B: "(...) Haben Sie dann auch gleich den Vorschlag für die Auswertung, wobei ich hier sagen muss, ist also meines Erachtens nach sind diese Vorschläge in der Regel eher sinnlos. Das geht vom Lagetyp los, dann der Rhythmus und häufig werden die Pegel nicht erkannt. ST-Hebungsinfarkte denke ich mal eher ja, aber man muss das alles kontrollieren, so dass man sich hier auf diese automatische EKG-Analyse, auf den Vorschlag zumindest, vielleicht 5 Prozent orientieren kann. Alles andere ist falsch. Sodass wir das hier generell ausgeklinkt haben in unserer Klinik, dass gar keine Vorschläge erscheinen und wir also das EKG im Computer selber auswählen müssen und dort die Bausteine einarbeiten in die Auswertung."	Aktuelle Probleme bei existierenden automatischen EKG Auswertungen	Problem
Experte A: "Jeder, weil es eine solche Basis Untersuchung ist und die so viel wichtige und dringende Informationen enthält, dass man das nicht verpassen sollte." ***Experte B****:* "(...) die für mich jedenfalls seit Jahrzehnten wichtigste nicht invasive Funktionsdiagnostik.(...) Ohne ein EKG registriert zu haben bei Verdacht auf Myokardinfarkt, wäre eine grobe Pflichtverletzung."	Wichtigkeit von EKGs	Vorgehen mit EKGs
Experte B: "(...) wo ich ohne EKG nicht hinkomme, sind Rhythmusstörungen, Hypertrophie einzelner Teile, Vorhof Ventrikel Hypertrophie, Schenkelblöcke und natürlich Rhythmusstörungen und Myokardinfarkt."	Diagnose von Krankheiten durch EKG	Anwend-ungsfall
Interviewer*:* "Gibt es denn Alternativen zu EKGs?" ***Experte B:*** "Nein. Es gibt ergänzende diagnostische Methoden, sprich also dann die die Echokardiografie. Wenn ich dann den Verdacht habe auf einen Myokardinfarkt beispielsweise, (...)."	Alternativen zu EKGs	Vorgehen mit EKGs
Experte A: "Aber das ist nicht die einzige Gelegenheit, wenn. Das ist jetzt hier im Krankenhaus so. Beim Hausarzt ist es so, dass wenn ein Patient/ eine Patientin Beschwerden vorträgt, die auf eine Herzerkrankung deuten könnten, dann wird sicher ein EKG gemacht. "	Anwendung von EKGs beim Hausarzt	Anwend-ungsfall
Experte A: "Die meisten Hausärzte können das auch und sollten das eigentlich haben."	Fähigkeit und Verfügbarkeit der EKG Durchführung beim Hausarzt	Vorgehen mit EKGs
Experte A: "In anderen Disziplinen ist das nicht so. In der Unfallchirurgie, wenn jemand jung sich den Knöchel bricht, dann braucht er kein EKG."	Keine Anwendungs-fälle von EKGs	Anwend-ungsfall
Experte A: "Das macht die Triage Schwester. Sie macht es aber gut und richtig. Und wenn sie sagt, das ist ein Patient für die Innere Medizin, dann soll dieser Patient/ diese Patientin ein EKG bekommen. Bekommt er auch immer."	Zuordnung eines Patienten zur Inneren Medizin	Vorgehen mit EKGs
Experte A: "Im wesentlichen Beschwerden, die entweder auf eine schwerwiegende Rhythmus Störung deuten oder aber für einen Herzinfarkt typisch sein können. Und da der Herzinfarkt so viele verschiedene Erscheinungsformen hat, setzt man die Schwelle ziemlich niedrig an."	Hinweisende Beschwerden für eine Herz-erkrankung	Vorgehen mit EKGs
Experte A: "Ich möchte, dass die Ärzte und Ärztinnen die EKGs angucken, das besser können. Ja, das finden die selber auch. Also gerade Studenten im letzten Abschnitt der Ausbildung und Ärztinnen, die beginnen mit ihrer Berufsausübung,	Problem von mangelndem	Problem

empfinden selber diesen Mangel, dass sie das zu wenig trainiert haben. Und der ist auch real dieser Mangel. Merken wir auch selber. Und wir setzen alles daran, noch mehr Ausbildung, noch mehr Ausbildung, noch mehr Ausbildung zu machen, damit Ärztinnen und Ärzte besser EKGs lesen können."	Wissen vieler Ärzte über EKG Auswertung	
Experte B: *"*Ich würde verbessern wollen: die schnellere Handhabbarkeit, die bessere EKG Aufzeichnung und was ich für meinen Teil empfinde, es müssten mehr EKG Ableitungen als routinemäßig 12 registriert werden. Ich brauche also im Prinzip, das wird sicher bei anderen anders sein, mehr Ableitungen als diese ursprünglichen 12. (...) ich muss wissen, was hat der Patient klinisch für Daten, Anamnese, Status und würde dann verlangen EKG und teilweise auch erst während der EKG Schreibung entscheiden. (...). Also man kann nicht sagen routinemäßig mehr als zwölf, aber eben nicht nur. Das ist das Minimalprogramm die 12."	Problem von Handhab-barkeit, Aufzeichnung und geringe Anzahl an Kanälen/Ableit ungen	Problem
Experte A: "Ich würde für jeden, der in der Inneren Medizin ausgebildet wird, das sind Allgemeinärzte oder aber Internistinnen/ Internisten, möchte ich, dass die das können. Ob ein Unfallchirurgen oder ein Orthopäde oder ein Neurochirurg, der muss das nicht so gut können. Oder ein Psychiater. Aber klar, der Allgemeinarzt/Internist, die sollen das können. (...) Zum Staatsexamen kann man Grundkenntnisse voraussetzen. Wenn man das nicht übt, verblassen die sofort."	Voraus-gesetztes EKG-Wissen bei verschiedenen Arztgruppen	Vorgehen mit EKGs
Interviewer: *"*Wie lange würden Sie dann, würden Sie brauchen, um ein EKG jetzt in eines oder mehrere dieser Fünf Klassen einzuordnen?" ***Experte A:*** "Um zu sagen, dieses EKG ist normal, brauche ich nicht länger als 12 Sekunden. Okay, wenn ich sage, da stimmt was nicht, dann kann es auch noch sehr schnell gehen. Manchmal diskutieren wir aber zu mehreren auch 5 Minuten über einen EKG." ***Experte B:*** *"*Ein paar Minuten denke ich mal, denn wir werten ja am Tag hunderte von EKGs aus. Wir haben einzelne Kliniken zugeordnet, für die wir die EKG auswerten."	EKG Auswertungs-zeit von Ärzten	Vorgehen mit EKGs

Usability Testing von EKG Kontrafakten		
Textabschnitt	**Kodierung**	**Kategorie**
Query Nr: 394, Query Label: 4, Query Prediction: 4, Query ExperteA: 4, Query ExperteB: 4, NG Nr: 627, NG Label: 3, NG Prediction: 3, NG ExperteA: 3, NG ExperteB: 3		
Experte A: "Also kann ich jetzt schon sagen, dass es nicht normal ist, die einzige Frage ist, ob hier eine Erhebung ist und dazu fehlt mir die Eichung, weil es jetzt auf die echte Maßzahl ankommt. Würde man denken Ja, das ist ein Herzinfarkt, aber kein ganz frischer. Aber die Hebung, die dazugehört, ist nicht ausreichend und ich würde das der Klasse 4 zuordnen." ***Experte B:*** *"*Hier würde mir die T-Wellen Negativierung auffallen bei V2, V1, Bis V4, das wäre dann hier die Nr. 4."	Einordnung in eine Klasse	Vorgehen mit EKGs
Experte A: "Na ja, das Rote EKG ist für mich ein anderes EKG und deshalb befunde ich es anders. Und mehr kann ich dazu nicht sagen. Jetzt ist die Frage also spannend. Wäre ja gewesen, um wieviel sich diese Teile verändern müssten bis ich sage jetzt ist es normal. Hier ist es so stark verändert, dass ein Blick klar ist, dass es normal ist. (...) das ist eigentlich ein Kontrast hier das Rote und der Kontrast ist so stark, dass es völlig klar ist Das ist jetzt normal. Wenn Sie Schwellen rauskriegen wollen, ab wann ich das finde, dann ist das zu stark der Kontrast. (...) das, was hier verändert wurde, weiß dieses EKG ganz eindeutig einer anderen Klasse zu." ***Experte B:*** *"*Richtig. Also hier stellt sich das R positiv auf. Ja, ich würde also hier auch im Vergleich die T-Wellen Abflachung gewesen war, im anderen was wir hatten. Hier ist die Abflachung. Das wäre aber noch physiologisch. Und hier haben sie eingezeichnet dieses Rote, dass die T-Wellen also insgesamt über den gesamten Brustwand. Hier kann es negativ sein, aber hier ab V2 dann schon nboth experts agreed that the counterfactuals look like real ECG time series.icht mehr. Und hier wird es also ab bis V6 hin immer weiter positiv. Also das finde ich sehr gut. Das ist also dann zur Normalisierung hin. Hier würde man sagen unauffällig normales EKG. (...) Das ist also sehr gut gemacht hier wirklich sehr gut übertragen mit überlagert diese Kurven. (...). Auch hier ist wieder diese Normalisierungs-Kurve muss ich sagen, sehr gut dargestellt, in dem Kanal 3 besonders. Also hier geht es aber um die Amplitude, nicht um die T-Wellen."	Meinung zu Kontrafakt	Kontrafakt
Interviewer: *"*Könnten Sie da jetzt selber so ein Kontrafakt einzeichnen? (..)" ***Experte A:*** "Es soll dann sicher normal sein? Dann ist das, was gemacht wurde, schon ganz richtig. Also ich brauche dann hier eine isoelektrische und dann eine positive T Welle." ***Experte B:*** *"*Diesen Kammerkomplex muss ich hier einzeichnen, dass der nur so ist und auch hier. (...) hier dann deutlich positiver. Die müssen positiver sein. Das würde mir also auffallen vom Pathologischen zum Physiologischen hin."	Selbst-entworfener Kontrafakt	Kontrafakt
Query Nr: 921, Query Label: 0&2, Query Prediction: 0&2, Query ExperteA: 0, Query ExperteB: 0, NG Nr: 365, NG Label: 3, NG Prediction: 3, NG ExperteA: 3, NG ExperteB: 3		
Experte A: "Ich würde es der Klasse Leitungsstörung zuordnen. Also Klasse 0." ***Experte B:*** *"*Erregungsbildungsstörung denke ich mal. Die null."	Einordnung in eine Klasse	Vorgehen mit EKGs
Experte A: *"*Das Problem ist, dass mir die Eichung hier fehlt. Zum Schluss hängt es von Zahlen ab. Also wenn ich mich damit beschäftige und ich brauche sowohl die Eichung und das ist keine richtige Eichbaum Zacke, ich brauche eine Eichbaum Zacke, und ich brauche die Angabe, was wieviel 100 Millisekunden sind. Und dann, ganz zum Schluss kann ich dann nicht entscheiden, dass wir das jetzt hier wirklich so mit Eyeballing machen."	Kritik an EKG Darstellungs-art	Design Anforder-ungen

Experte A: "(...) diese tiefe Zacke hier gehört da nicht hin, und die führt mich mit den Augen sofort auf die Ableitung V1 und in der Ableitung V1 ist die Zeit von hier bis hier eindeutig zu lang. Und das ist eine Leitungsverzögerung. (...) Diese Verzögerung macht einen Rechtsschenkelblock daraus und je nachdem wie breit das jetzt ist, wäre es ein kompletter oder inkompletter Rechtsschenkelblock. (...). Ich glaube, dass es auch zusätzlich ein AV Block ersten Grades ist und außerdem ist es ein überdrehter Links Typ. Und dann haben wir also einen Bifaszikulären Block Anthenuanen Typs plus AV Block ersten Grades. alles gehört in die Klasse Leitungsstörung. Und wenn dieser Patient mir sagt, er sei schon mal bewusstlos geworden, dann braucht er einen Schrittmacher, und das ist jetzt nicht banal." ***Experte B:*** "Also das ist ein Rechtsschenkelblock, denke ich mal bei der Frequenz. (...) weil wir hier einen überdrehten Linkstyp. Ich sehe das praktisch hier in der Abteilung V1 , dass ich hier eine R Zacke habe und die Nachschwankung ist im Prinzip hier diskordant zum Kammerhauptausschlag. Normalerweise wäre ja der normale Kammerkomplex in V1 so schmal und gegenläufig und der hier ist zwar auch gegenläufig, aber breit. Und hier haben wir sogar diese "m" Form für den Rechtsschenkelblock. (...) Dann wieder dieses Runde S, was in Ableitung 1 und V6 ähnlich sein sollte für die Diagnose mit Schenkelblock."	Medizinische Erklärung der Experten Einordnung	Vorgehen mit EKGs
Interviewer: "Die KI hat auch Klasse 0, Conduction Disturbance vorhergesagt. und auch Myokardinfarkt." ***Experte A:*** "Das muss man ablehnen. Da muss man das System nachschulen. mit dem Herzinfarkt." ***Interviewer:*** "Es ist eigentlich richtig vorhergesagt worden. Das heißt, die Ärzte, die hier die Klasse zugewiesen haben, haben auch gesagt, dass es ein Herzinfarkt ist. ***Experte A:*** (...) Hier ist kein Herzinfarkt zu sehen. Das Problem ist das, was man für den Herzinfarkt braucht, sind wie eben in dem EKG diese Strecken. Und beim kompletten Rechtsschenkelblock sind diese Strecken verändert. Das weiß man aber eigentlich. Das lehne ich ab." --- ***Experte B:*** "Für einen Infarkt dürfte hier kein Q sein. Hier nicht und in Ableitung AVF auch nicht. Und auch hier sehen Sie es als Minimales hier dieses ganz Kleine. Das wäre entscheidend, dass es keinen Myokardinfarkt ist. Hier dieses kleine R. (...) Das bekommen wir sehr häufig hier in der Auswertung. Wenn dort irgendwelche Schenkelblöcke sind, dass dort einen Infarkt vorhergesagt wird. Das ist aber mit Sicherheit falsch. Beim gerade beim Rechtsschenkelblock kann ich also Überlagerungsfrei sowohl Vorderwand- wie Hinterwandinfarkte diagnostizieren."	Kritik an KI Vorhersage	KI
Interviewer: "Hier wieder der Kontrafakt. (...) würden sie auch den Kontrafakt der ganz normalen Klasse zuordnen und für wie sinnvoll erachten Sie diesen?" ***Experte A:*** "Ja, das ist ein normales EKG. Jetzt sind alle diese Phänomene weg, also der QRS komplex ist wieder schmal, die Leistungsverzögerung hier ist weg und der Lagetyp ist jetzt wieder normal. Und die PQ Zeit ist auch geringer geworden. Ja, das ist eindeutig normal. Aber wie gesagt, die Schwellen. Bei diesen Leitungsverzögerungen geht es um Messwerte, um Millisekunden, und da muss man ein Lineal nehmen. Das ist jetzt wirklich grob." ***Experte B:*** " Ja, hier sieht man also wieder eindeutig den schmalen Kammer Komplex im Vergleich zu beiden. Hier wäre also auch eine positive T Welle erforderlich, ähnlich wie hier auch schon drin ist. Und hier wäre dann ja auch der positive Ausschlag positive Ausschlag. (...). Ja, ist auch drin hier die das kleine R und hier ähnlich wie ich es ja schon eingezeichnet habe, in V1 nicht ganz so tief gezeichnet, aber das wäre im Prinzip jetzt hier das Normale auch in V2 überlagert. Und auch hier keine breite S Zacke, sondern praktisch dieser schmale Kammerkomplex, wo dann unmittelbar nach dem R das S noch kommt und dann hier auch eine nicht ganz so ausgeprägte T Positivität. Also da wäre ich auch sehr einverstanden mit dieser Überlagerungs-Kurve zum Normalen, also vom Rechtsschenkelblockbild zum normalen EKG."	Meinung über Kontrafakt	Kontrafakt
Interviewer: "Können Sie den Kontrafakt so einzeichnen, sodass es wieder ein normaler Herzschlag wird." ***Experte A:*** "Es waren viele Stellen, die man verändern muss, damit es normal wird. (...) diese tiefe S-Zacke gehört weg. Diese Tiefe gehört weg, die darf bleiben. Und diese Tiefe muss weg. Und hier muss es so aussehen und schmaler. Beides muss weg, das muss weg, das muss, das muss weg. Und es muss das P näher an das QRS ranrücken und wenn das der Fall wäre, dann würden auch alle diese Endteilveränderung verschwinden. Also das ist alles Pathologisches, gehört aber zu dieser Leistungsverzögerung dazu. Das, was das System fälschlicherweise als Herzinfarkt bezeichnet hat." ***Experte B:*** "Hier unten müsste das ja noch ganz schmal sein. Der Kammerkomplex ist mir nicht so gelungen, aber ganz schmal."	Selbst-entworfener Kontrafakt	Kontrafakt
Query Nr: 1212, Query Label: 3, Query Prediction: 3, Query ExperteA: 1oder3, NG Nr: 425, NG Label: 0&2, NG Prediction: 0&2, NG ExperteA: 3		
Experte A: "Also da fehlt mir jetzt wieder die Eichung, weil ich wissen muss, wie hoch ist diese Zacke und wie ausgelegt ist diese Zacke Ich nehme mal an, dass das so ist, dass es einen bestimmten Grenzwert überschreitet, sodass ich das der Hypertrophie zuordnen würde. Eins aber das hängt also ohne Lineal, ohne Eichkurve ist es schwierig, ansonsten wäre es normal. (...) Ansonsten ist es ein gesunder junger Mann, der langsam ist bzw. der Sport macht. Aber ja, das hängt an den Eichzacken. Eyeballing und wenn ich nicht messen kann, dann ist es so."	Einordnung in eine Klasse	Vorgehen mit EKGs
Interviewer: "Ich habe hier dazu wieder den Kontrafakt und dieses Mal, wie sich der normale Herzschlag verändern müsste, dass es krank werden würde. (...) In diesem Fall was das System gefunden hat wäre Conduction Disturbance als auch Myokardinfarkt." ***Experte A:*** "Also hier ist definitiv kein Herzinfarkt zu sehen. Können wir ausschließen. Eine Leitungsverzögerung, dann müsste man messen können, wie weit das hier ist. Nein, eigentlich ist das auch für mich ein normales EKG. also dem würde ich widersprechen. Die sehen aber, dass diese Zacke kleiner geworden ist. Da würde ich jetzt auch die Vermutung für einen Hypertonus nicht mehr haben. Aber da das schon als normal galt, ist das auch normal. Aber Sie sehen, dass diese Zacke kleiner geworden ist und ich in Kombination mit der Zacke nicht mehr glauben würde, dass ein Hypertonus vorliegt. Also eher normaler gemacht."	Kritik an Kontrafakt	Kontrafakt
Query Nr: 474, Query Label: 2, Query Prediction: 2, Query ExperteA: 1oder3, NG Nr: 221, NG Label: 3, NG Prediction: 3, NG ExperteA: 3		

Experte A: "Also ich würde ganz gerne jetzt auch wieder messen, da gehts wieder um diese also um die Höhe dieser Zacke zusammen mit dieser Zacke, das kann ich nicht sagen. Also entweder das Normal oder es ist auch Hypertrophie und wir haben hier, es ist also minimal, die B Zacke ist ein bisschen erhöht, aber die hier nicht." ***Interviewer:*** "Mir ist die Wellenform hier aufgefallen. Also dass es sich hier so verändert. Hat das eine Bedeutung?" ***Experte A:*** "Nein, da hat er sich bewegt. Das ist völlig egal. Also diese Extremität hat sich bewegt und das ist der linke Arm, muss er sein. Und der hat ihn bewegt und da guckt man drüber weg. Das war genau so wie hier. Es ist belanglos."	Einordnung in eine Klasse	Vorgehen mit EKGs
Interviewer: "Vorhergesagt wurde ein Herzinfarkt." ***Experte A:*** "Hmm, ja, da ist hier das System zu empfindlich eingestellt. Ich weiß nicht, welche Kriterien. Ich nehme mal an, dass das System diese ST Strecken Hebung automatisch als Herzinfarkt erkennt. Mehrfach schon getan. Das ist machen unsere Anfänger auch alle. Aber es ist kein Herzinfarkt. Das alles ist eine angehobenen ST-Strecke. Die könnte beweisend für einen Herzinfarkt sein. Ist sie aber nicht, weil sie falsch rumgebogen ist. Also ihr System glaubt immer dann, wenn die ST Strecke gebogen ist ist es ein Herzinfarkt. Das ist aber nicht der Fall, sondern es kommt immer noch darauf an, wie rum. Solange die so gebogen ist, ist es kein Herzinfarkt. Wäre sie gehoben und so rumgebogen, wäre es ein Herzinfarkt, müssten Sie dem System noch beibringen."	Vor-geschlagener Grund für fehlerhafte KI Entscheidung	KI
Interviewer: "Dann hier wieder der Kontrafakt dazu, der versucht aus einem Herzinfarkt einen normalen Herzschlag zu machen. Würden Sie den pinken Graphen normal einordnen und finden Sie das sinnvoll?" ***Experte A:*** "Also er hat die Linien glattgezogen. Das war aber schon vorher nicht schlimm und er hat die ST Stecken hier abgesenkt. Und dann ist es für das System wieder normal, es war vorher schon normal."	Kritik an Kontrafakt	Kontrafakt
Query Nr: 1751, Query Label: 0, Query Prediction: 0, Query ExperteB: 0, NG Nr: 6, NG Label: 3, NG Prediction: 3, NG ExperteB: 3		
Experte B: "Das ist eine Rhythmusstörung. Wir haben wir praktisch eine absolute Arrhythmie, die Abstände sind völlig irregulär. Würde ich der Klasse Null zuordnen." ***Interviewer:*** "Mir ist die Wellenform hier aufgefallen. Also dass es sich hier so verändert. Hat das eine Bedeutung?" ***Experte B:*** "Hier wird werden entweder die Elektroden lose, das heißt wir haben hier drin keine Wechselstromüberlagerung, (..) oder die Elektroden sind trocken oder wir haben zusätzlich noch hier diese größeren Schwankungen innerhalb dieser Nulllinie. Das sieht man häufig auch bei atemabhängigen Schwankungen weiß man nicht. (...) Das wäre im Rahmen einer schlechten EKG Schreibung. Da kann man also nichts draus machen. Würde ich so erstmal sehen. Aber hier diese praktisch Arrhythmie, die fällt mir hier deutlich auf."	Einordnung in eine Klasse	Vorgehen mit EKGs
Experte B: "Ich weiß nicht, ob das so in den Automatikauswertung auch bei Ihnen eine Rolle spielt, dass dieser Algorithmus natürlich auch jetzt altersbezogen gesehen werden muss. Ich habe natürlich jetzt ein EKG, wenn der Patient drei Tage alt ist oder ob er 83 ist. Also da gibt es große Unterschiede und das vermisse ich auch so ein bisschen in der in der EKG Analyse, dass man also dort auch anbietet. Man gibt ja das Geburtsdatum nicht umsonst ein, also sollte man annehmen ein, dass das also praktisch angeboten wird. "	Vorschlag zur Erweiterung der KI Vorhersage mit dem Faktor Alter	Verbesser-ungsvor-schlag
Experte B: "Ja, hier müsste man jetzt sehen mit dem Zirkel. Also erst mal natürlich die rhythmischen Abstände, das hat man hier korrigiert, so dass man das völlig rhythmisch, ja wohl nicht mehr übereinander, sondern nebeneinander, weil man dort aus der totalen Arrhythmik in die Rhythmik hineinkommt. Das ist okay und auch die Kammerkomplexe, die waren ja nicht wesentlich verändert. Und jetzt ein normaler Typ geworden aus dem links Typ der primär war jetzt zum Normalbefund hin. Das wäre ja der normale Typ. (...) Okay, das ist also hier in Ordnung, vor allem sozusagen der normale S zu R Übergang von V1 bis V6, die normale Nachschwankung, Veränderung der positiven T-Wellen, keine S-T-Senkung. Das ist also hier sehr gelungen."	Zustimmung mit Kontrafakt	Kontrafakt
Interviewer: "Was ist hier mit dem AVL? Das sieht ein bisschen chaotisch aus." ***Experte B:*** "Kann aber so sein. (...) Das ist ja technisch gestört, ist also hochgradig gestört. Das EKG ist in 1 schon gestört, in 3 in Ableitung AVL auch noch mal. (...) Da sind also Störung und das wird natürlich auch dann die Normalisierungskurve nicht störungsfrei darstellen können. Sag ich mal das sind Artefakte, die im Rahmen jetzt dieses schon primär Gestörten EKGs zwar versuchen das zu normalisieren, aber auch dort sind Grenzen würde ich so sehen. (...) Man kann also nie sagen, dass man EKG immer vernünftig und störungsfrei ableiten kann. Entweder liegt an den Elektroden oder es liegt an der an der Trockenheit der Haut. Oder es liegt beispielsweise das haben wir natürlich sehr häufig bei Brust Elektroden, bei männlichen Patienten ein Fell haben (...), bei den Frauen ist es natürlich die Körperfülle. Gerade im Bereich der Brust muss man die Elektroden also so zu platzieren, dass fast störungsfreie EKGs erhalten werden. Um das Ganze zu umgehen, wird sehr häufig die Filter Taste gedrückt. Da sich also eine Glättung der EKG Kurve, dass dort die Störungen zwar minimiert werden, aber dann natürlich auch sämtliche Aufsplittung, die ich im richtigen EKG ohne Glättung der EIchzacke hätte, dass sie dann natürlich auch weg sind. Und das muss ich wissen."	Störung des EKGs und Kontrafaktes	Kontrafakt
Experte B: "Wird natürlich ein bisschen schwierig, dass man dort mit dem Zirkel arbeitet. Das ist schwierig, wenn man das macht. (...) wichtig wäre, dass hier die normalen RR Abstände wieder klarkommen. Also hier würde ich bei der Ableitung auf AVL würde ich nichts normalisieren, das würde ich so lassen wie es ist."	Selbst-entworfener Kontrafakt	Kontrafakt
Query Nr: 304, Query Label: 1, Query Prediction: 0, Query ExperteB: 4, NG Nr: 641, NG Label: 3, NG Prediction: 3, NG ExperteB: 3		
Experte B: "Also hier haben wir erstmal den normalen Rhythmus. Also hier würde ich sagen, müsste man das. S-T Veränderungen schauen, also ich würde das hier im Rahmen der 4. Ich schwanke zwischen Myokardinfarkt, aber das ist nicht ganz typisch. Aufgrund des Kammerkomplexes. Ich würde das eher aufgrund der Veränderung der S-T Strecke und der T Welle gerade bei der Brustwand Klasse 4 zuordnen."	Einordnung in eine Klasse	Vorgehen mit EKGs

Interviewer: *"*Die KI hat hier Klasse 0 vorhergesagt, was aber falsch ist. Tatsächlich. Also was hat der Arzt der 90er Jahren mal gesagt hat? War die Hypertrophie." *Experte B: "*Die Hypertrophie würde ich hier überhaupt nicht ins Auge fassen, weil zum ersten der Lage Typ dazu nicht passt. Hier würde ich sagen sind deutliche Veränderung der T-Wellen. Also die 4 würde ich mich dafür entscheiden. Deutliche Veränderung der S-T-Strecke und der T-Welle. allerhöchstens wenn man den Kamer komplexes ausmessen würde. Von der Breite her. Ob der jetzt schmaler sein müsste. Aber das kann man natürlich jetzt bei dieser Schreibgeschwindigkeit nicht sagen. Da müsste man also genau wissen, wie ist die Schreibgeschwindigkeit. Der Kammerkomplex ist schon etwas breit. Aber selbst wenn der breit wäre, könnte man auch hier wieder nicht sagen, Ist es jetzt das typische Links oder das typische rechts Schenkel Blockbild, das ist hier nicht einordnungsbar."	Diskussion über KI Entscheidung	KI
Experte B*: "*Ja, ja, das ist korrekt. Das ist korrekt. Hier wird er jetzt links. Typisch. Aber eben auch wieder diese Verzitterungen. (...) Es stimmt, dass der rote Graph ein normales EKG zeichnet."	Zustimmung mit Kontrafakt	Kontrafakt
Experte B: "Also runter bis V6 dann die T positiveren."	Selbst-entworfener Kontrafakt	Kontrafakt

Interview		
Textabschnitt	**Kodierung**	**Kategorie**
Experte A*:* "Also ich habe nichts gegen eine künstliche Intelligenz, die schlauer ist als ein gut ausgebildeter Arzt. Ich habe sie noch nicht gefunden (...) das mit der künstlichen Intelligenz müssen andere machen, weil das zu kompliziert ist." ***Experte B:*** " Ich denke mal, dass es recht gut gelöst."	Meinung zu KI	KI
Experte A: "Ich würde meine Anstrengungen weiter darauf verwenden, Ärztinnen und Ärzte so gut auszubilden, dass sie möglichst fehlerfrei EKG lesen können."	Zukünftiges Vorgehen zur EKG Klassifizierung	Vorgehen mit EKGs
Experte A: "Aber ich warne bei der Ausbildung alle davor, diese automatisierten Auswertungen zu genau anzugucken. (...) ST Strecken Veränderung bleibt im Moment offenbar ein wenig dem kundigen Arzt überlassen. Da sind bei den Auswertungen jetzt noch Fehler drin und das muss man sich selbst angucken."	Aktuelle Probleme bei existierenden automatischen EKG Auswertungen	Problem
Experte A: "Wir haben ja schon bei allen EKG Schreibungen eine automatisierte Auswertung drin und die ist gut für die Messung der Zeiten, da spare ich mir Zeit. Wie breit jetzt ein QRS Komplex ist, das zeigt mir das Gerät, das machen die auch immer richtig, sodass ich da mal Lineal nicht rausholen brauche oder für die PQ Zeit. das macht das Gerät richtig. "	Anwendungs-möglichkeiten der KI zur Unterstützung bei EKG Auslesung	Anwend-ungsfall
Interviewer:" Okay, was halten Sie jetzt von dieser kontrafaktischen Erklärungsmethode? Haben Sie davor schon mal davon gehört?" ***Experte A:*** "Also die Idee ist super, eigentlich. (...) Aber ihnen geht es ja mehr um die Methode. Mit diesem Kontrafakt. Und da würde ich mal sagen: der ist gut. Ja vielleicht sogar für echte Menschen was zu lernen und für die KI bestimmt. " ***Experte B:*** "(...) Das gibt mehr Information, wenn ich diese Überlagerungskurve sage ich mal mit dazu noch hätte. Also ich fand es sehr interessant und auch sehr aufschlussreich. Und gerade für jüngere Kolleginnen sage ich mal, dass die also den Vorschlag haben. So müsste das also im Normalfall aussehen und nicht alleingelassen werden mit der Interpretation der pathologischen Kurve. Die sind häufig überfordert. Da finde ich das schon sehr gut. Diese Normalkurve mit anzuzeigen oder zumindest den Vorschlag zu machen, so müsste das im Normalfall aussehen, finde ich an sich eine sehr gute Idee. Also die Methode insgesamt gefällt mir ganz gut."	Meinung zu Kontrafakten	Kontrafakt
Experte B: "Verbesserungswürdig wäre natürlich hierzu von den EKG-Kurven her, dass man wüsste, die Eichung und das man wüsste von der Breite her. Das würde sicher dann noch mehr Sicherheit bringen. Auch in der EKG Auswertung."	Verbesserungs-vorschläge des Designs	Verbesser-ungsvor-schlag
Experte B: "In Deutschland ist der Standard 1 cm gleich 1mV von der Amplitude her und 50 Millimeter pro Sekunde Papier Forschung und nur in Einzelfällen 25 im angloamerikanischen Sprachraum sind also 25. Wird in vielen ambulanten Praxen auch gemacht. Aber nicht weil's der Standard ist in den anderen Ländern, sondern um Papier zu sparen, muss man schon so sagen."	Design Standard von EKGs	Vorgehen mit EKGs
Experte B: "Also er muss zumindest das aufzeigen, was von der Pathologie jetzt wieder in die normalen Befunde gehen könnte. Da ist natürlich die Schwierigkeit, also verschiedene, von diesen Klassifizierungen angefangen, von den Abläufen dann zu erkennen. (...) Gerade bei den Schenkelblockveränderungen und Nachschwankungen sekundär bedingt. Die müssen wieder in die normalen Erregungsrückbildungszustände hineingehen. Das wäre das wichtigste Ziel oder wichtigste Aufgabe für die "Normalisierung" der Kurve. Weil dort natürlich die Differenzierung hinsichtlich des Schwierigkeitsgrades dieser Normalisierungskurven enorm auseinander weicht."	Sinnhaftigkeit eines Kontrafaktes	Kontrafakt

Experte A: *"*Sie müssen die natürlich granulieren. Sie müssen Stufen einführen und sagen Ich verändere ganz wenig, dann etwas mehr, etwas mehr. In Inkrementen sowas machen. Und sagen so, das ist die Veränderung, die das System bewogen hat, das jetzt so oder so zu klassifizieren. Und diese Veränderung führe ich in winzigen Schritten zurück. Und ab wann ist es dann normal. Er muss ganz fein granulär arbeiten. (...) An den Stellen auf, die sie jetzt auch der KI beibringen wollen."	Verbesserungsvorschläge eines Kontrafaktes	Verbesse-rungsvor-schlag
Experte A*:* "(...) Leitungsverzögerungen braucht man das nicht zu wissen, weil es eben misst, dann entscheidet sich das an den Messwerten. (...) Da brauchen Sie die Methode nicht, weil das hängt an den Messwerten und da muss man nichts dazu sagen."	Klassen mit schlechter Anwend-barkeit von Kontrafakten	Anwend-ungsfall
Experte A*: "*Bei den ST Strecken ist es total knifflig, weil die Frage ist, ab wann klassifiziert das System, das wirklich als pathologisch und da ist auch im Angucken ein ganz bisschen Korridor drin. Und insofern dafür ist es interessant, aber dann wirklich mit winzigen Inkrementen."	Hilfreiche und gute Anwend-barkeit von Kontrafakten	Anwend-ungsfall
Experte A: "(...) Da wäre es also das Wichtige, dass man die als pathologisch klassifiziert und sagt: die sind so pathologisch, dass ich die Einstellung des Ärzten kennen muss. (...) Das System sollte ein normales EKG erkennen. Vielleicht ist das das überhaupt das Allerbeste, dass man erst mal sagt, dass das, was wir in 12 Sekunden können, nämlich dass es normal ist. Das bringen Sie dem System jetzt mal zuerst bei und dann sollten solche fälschlichen Myokardinfarkte dann eigentlich mal raus. (...) das System sollte auf jeden Fall ein normales EKG sicher als normal befinden mit allen Spielarten, die es im Normalen gibt. Und dann sollte man bei den pathologischen EKGs, die man dem System zum Lernen gibt, die Kontrafakten tatsächlich in winzigen Schritten beibringen."	Anforderungen an eine KI zur EKG Klassifizierung	KI
Experte A: "Ich könnte mir das für den Unterricht ganz gut vorstellen. Also wenn man jetzt minimale Veränderungen einführt, dann könnte man die auch gut im Unterricht, weil es derjenige oder diejenige, die EKG lesen lernt, macht ja nichts anderes. Am Anfang sagt mir das gar nichts. Bis ein EKG anfängt zu sprechen. Und es gibt Dinge, die schreien einen an, und es gibt EKGs, da muss ich sagen, ich weiß nicht, da muss ich ganz genau hingucken und es ist total normal. Und dieser Lernvorgang hängt natürlich daran, dass man die kritischen Stellen in ihren Unterschieden richtig bewertet und sieht. Und deshalb könnte man solche minimalen Veränderungen gut als Unterrichtsmaterial verwenden." ***Experte B:*** *"*Das planen wir auch an, dass man also gerade den jüngeren Leuten in der Ausbildung das anbietet, sag ich mal als Möglichkeit, dass einem das nicht so schwer fällt in der Interpretation des EKGs."	Anwendung von EKG Kontrafakten im Unterricht	Anwend-ungsfall
Experte A: "Da fährt jetzt ein Notarzt im Dienst und ist Orthopäde und der holt einen Patienten mit Brustschmerzen ab. Und da schreibt er ein EKG und dann guckt er darauf, wie die Kuh vom Sonntag, er weiß davon nichts. Und es wäre schön, wenn die KI zuverlässig sagt, dass es ein normales EKG ist und man kann ihn wieder abliefern im Krankenhaus und sagen: mache dir keine Sorgen. Oder aber du transportierst hier gerade einen Patient mit einem akuten Herzinfarkt. Dazu müsste das Gerät noch mehr lernen. Das ist eine Anwendung, die ich total sinnvoll fände, weil ich werde das dem Orthopäden nicht mehr beibringen, er fährt aber trotzdem weiter Notarzt."	Anwendung von EKG Kontrafakten für Notärzte	Anwend-ungsfall
Interviewer: "Würden Sie sagen, dass sich ihr Vertrauen in eine Vorhersage einer KI erhöht mit Hilfe von solchen Kontrafakten?" ***Experte A:*** *"*Zurzeit noch nicht, mit dem was sie mir gezeigt haben." ***Experte B:*** *"*Ja, mit Kontrafakten würde ich das zumindest als besser erachten als ohne. "	Vertrauen durch Kontrafakten	Kontrafakt
Experte A: "(...) heute, ein gut ausgerüstetes Krankenhaus schreibt ein EKG in eine EDV-App hinein. Die gibt sich auch schon Mühe, das auszuwerten, aber zumindest sind die alle elektronisch gespeichert, so dass man auch die persönlichen Daten entfernen kann. Ganz leicht."	Zugang zu EKG Daten in Kranken-häusern	Vorgehen mit EKGs
Interviewer: "Sie haben ja auch die Wahrscheinlichkeiten gesehen, mit welcher sich das System sicher ist, dass es in diese Klasse gehört. Würde Ihnen sowas auch weiterhelfen und mehr Vertrauen in Vorhersagen schenken?" ***Experte A:*** *"*Naja, wir gewöhnen uns immer mehr an solche Wahrscheinlichkeiten und versuchen damit zu arbeiten. Wenn wir die Wahrscheinlichkeit nehmen, mit der ein erstbefundetes EKG hier im Haus richtig befunden wurde, dann bin ich mit 98 Prozent sehr zufrieden und auch beeindruckt, aber damit ich dieser KI Vertraue, hätte ich so 98% und größere Rahmen schon ganz gern. Nur wenn sich das System bei offenkundigen Fehlern so sicher ist, dass das stimmt, was es sagt, dann sehe ich der Wahrscheinlichkeit das nicht an, wie gut die KI ist." ***Experte B:*** "Ich würde denken eher nicht."	Nutzung und Kritik an Sicherheits-wahrscheinlich keiten der KI Vorhersagen	KI
Experte A: "Alles wo ich selber das Lineal brauche, um was zu messen, da kann sie mich gerne unterstützen, das finde ich super. Das mit dem Herzinfarkt ist eigentlich nicht schwierig, es haben aber viele damit Schwierigkeiten, und das System offenbar auch. Genau das ist besonders dringlich und besonders wichtig. Und was immer schwierig ist sind bestimmte Rhythmusstörungen, komplexe Rhythmusstörungen. Das könnte ich mir vorstellen, ist ein gutes Feld für eine KI. Dass es einen unterstützt. Dabei darf man auch in Alternativen denken: Wenn es das nicht ist, dann könnte es das sein. Was ist jetzt wahrscheinlicher. Das sind so EKGs an denen brüten wir ziemlich lange, um zu sagen, was für eine Art Rhythmusstörung das jetzt eigentlich ist. Und das ist jetzt aber nicht so akut wie der Herzinfarkt. Das muss ich wissen, weil der Patient muss, dann auf dem Kathetertisch. Bei bestimmten Rhythmusstörungen kann ich mir das nämlich auch morgen noch mal angucken und sagen, wie sieht es denn heute aus und das machen wir auch. Aber da wäre eine Unterstützung hilfreich. Aber das ist wirklich schwierig. Und ja, da finde ich Unterstützung fast hilfreicher." ***Experte A:*** "Sie werden also eher zum Erfolg kommen bei allen Sachen, die sowieso primär vom Messwert abhängen. Das mit der Hypertrophie ganz genauso, da braucht man einen Messwert, wie groß ist die Zacke. Dagegen die ST Strecke	Eignung der Nutzung einer KI für verschiedene Herz-krankheiten	Anwend-ungsfall

bereitet auch allen, die es lernen sollen, am meisten. Schwierigkeiten und komplexe Rhythmusstörungen hatten sie jetzt gar nicht dabei." ***Experte B:*** "Schwierig kann die Klassifikation von pavialen Speicherkrankheiten sein: up to . Da gibt es zwar auch Hinweise auf die Möglichkeit der EKG Befunde, aber es gibt dort keine mir bekannten spezifischen Befunde, wo man sagen könnte, aus diesen Oberflächen EKG heraus bleibt nur diese Diagnose einer kardialen Amyloidose übrig. Ansonsten natürlich die S-T.Hebungsinfarkte. Das ist alles abgenutzt. Sag ich mal, da gibt es nicht viel neues."		
Experte A: "Wir haben immer 12 Kanal EKGs und wir sind gewohnt, in Deutschland, die mit einer bestimmten Geschwindigkeit zu sehen: 50mm pro Sekunde und mit einer bestimmten Eichung. Das heißt 1 cm ist 1 mV. Und wenn man das gewohnt ist, dann will man die so sehen. In vielen anderen Ländern ist die Schreibgeschwindigkeit nur halb so schnell, die sind das gewohnt, die könnten mit unseren EKGs wenig anfangen. Also ich möchte das EKG so ausgedruckt sehen, wie ich es gewohnt bin, weil ich in meiner Art und Weise des Anguckens. (...) Ich sehe es eigentlich lieber das sechs Kanäle auf einer DinA4 Seite und die nächsten sechs, auf der nächsten. Dann ist es größer. Aber auch das ist eine Gewohnheitsfrage. Also mein Großvater ist Kardiologe gewesen in den 20er Jahren des letzten Jahrhunderts. Hat Drei Kanäle gehabt. Der fand das schon super? Da würden wir heute sagen es ist reinste Hellseherei. Wir ringen immer damit, ob wir mehr Ableitungen schreiben wollen, weil es Ableitungen gibt, auf denen was zu sehen ist für einen Herzinfarkt mit spezieller Anordnung, die auf diesen 12 Abbildung nicht drauf sind."	Genutzte EKG Aufzeichnungsart	Vorgehen mit EKGs
Experte B: "Ein Kanal ist abartig. Kann ich nicht nehmen, sondern ich brauche also zumindest schon die peripheren Ableitungen. (...) eher mehr als weniger. (...) Dann müsste man natürlich einen Kanal aussuchen, der Amplituden reich ist, also sowohl die P Welle, Kammerkomplex und auch die Nachschwankung. In den EKG Kochbüchern sag ich mal wird immer empfohlen zur manuellen Auswertung die Ableitung 2 der Standardableitung hinzuzuziehen. Das macht auch in zirka 90 Prozent gehen mitunter sehen sie aber in der V2 auch so mickrige Ausschläge, dass sie sagen ich muss eine andere Ableitung nehmen um die Messdaten zu erheben."	Erkenntnis-reichste Ableitung	Vorgehen mit EKGs
Interviewer: "Zeigen Sie, ob Ihre Patienten auch manchmal ihr EKG und erklären Ihnen, woran zu sehen ist, dass Sie krank sind?" ***Experte A:*** "Nein. Also wenn selbst die Ärzte, die wir ausbilden, Schwierigkeiten haben, das zu machen. Ich zeige den Patienten viel: Röntgenbilder von Blutgefäßen zum Beispiel, weil man das sofort erfassen kann. Beim EKG sage ich immer, das ist jetzt zu komplex. Da sage ich Ihnen nur, was darauf zu sehen ist." ***Experte B:*** "Nein, und zwar das aus gutem Grund, weil ich denke, dass durch eine EKG Überinformationen des Patienten mehr Patienten zu Tode gekommen sind als üblicherweise. Also da wäre ich sehr zurückhaltend. Ich bin kein großer Freund von Patienten mäßiger Demonstration der EKG-Kurven." *Interviewer:* "Man kann sich vorstellen, Ihnen das zusammen mit einem Kontrafakt zu zeigen?" ***Experte A:*** "Das lohnt ja nicht. Er muss mir das trotzdem glauben, dass er einen Herzinfarkt hat und das EKG ist ein extrem schlechtes Beispiel für den Patienten." ***Experte B:"*** Würde ich auch nicht. Würde ich auch generell nicht."	Erklärung des EKGs für Patienten	Vorgehen mit EKGs
Experte A: *"(...)* Aber jetzt noch mal hier mit den wo ich das eingemalt habe mit dem falschen Myokardinfarkt. Das male ich jeden Tag 5-mal auf, wenn ich meine Assistenten unterrichte, weil das machen die immer, wirklich immer falsch."	Probleme mit Herzinfarkt Klassifizierung	Problem
Experte A: " Das ist natürlich dann für jedes Lernen der KI ein Problem, wenn man ihr das schon falsch beibringt, dann kann sie es auch nicht Besonderes. Deshalb brauchen sie ganz oft möglicherweise Rückkopplung. "	Problem von falsch gelabelten Daten	Problem

Conspicuous Segments – Expert A

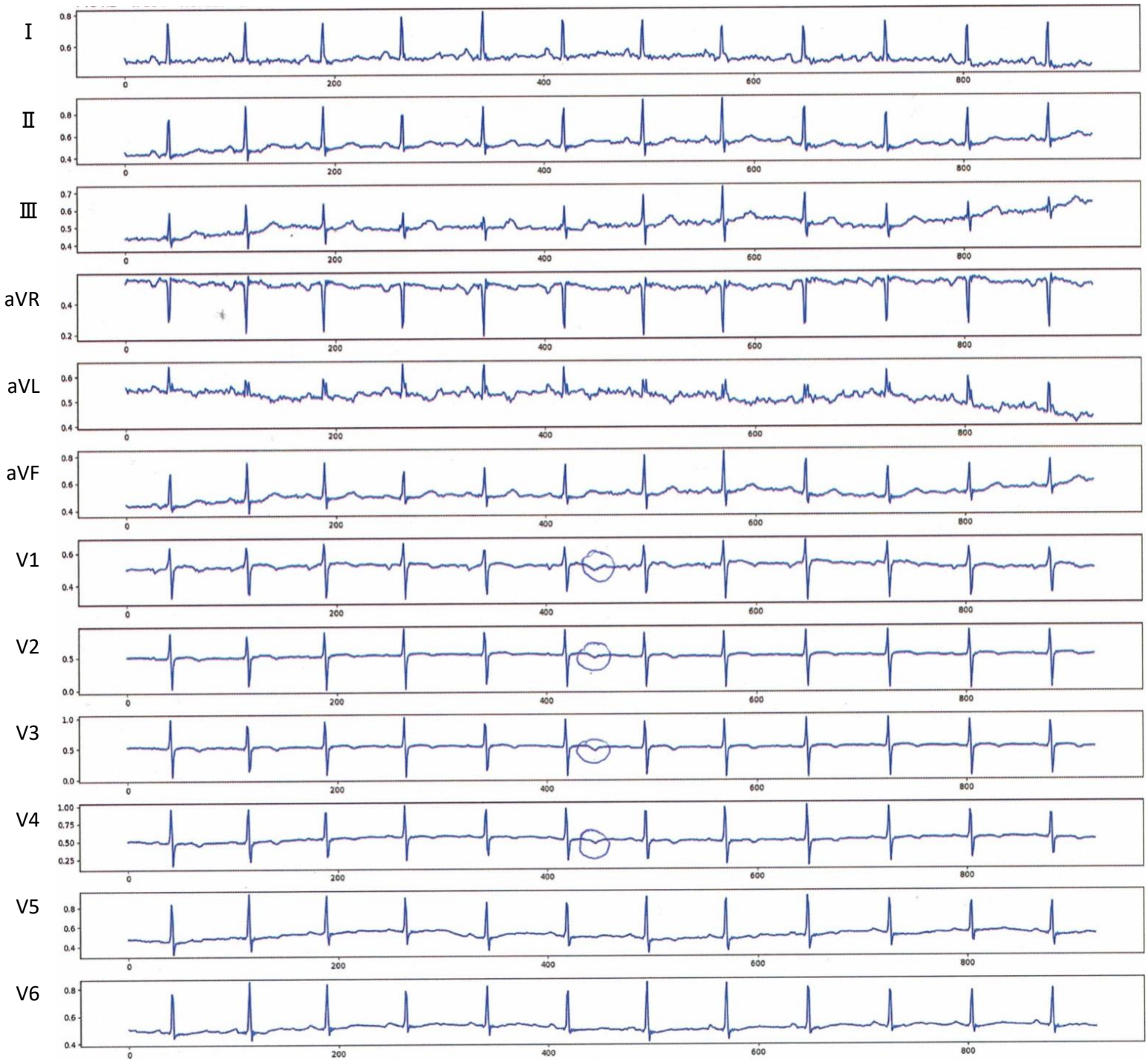
Expert Prediction: ST/T Change
Label: ST/T Change
AI Prediction: ST/T Change (94,82%)

Dataset: PTB-XL, Query: 394, NG: 627

Conspicuous Segments – Expert B

Expert Prediction: ST/T Change
Label: ST/T Change
AI Prediction: ST/T Change (94,82%)

Dataset: PTB-XL, Query: 394, NG: 627

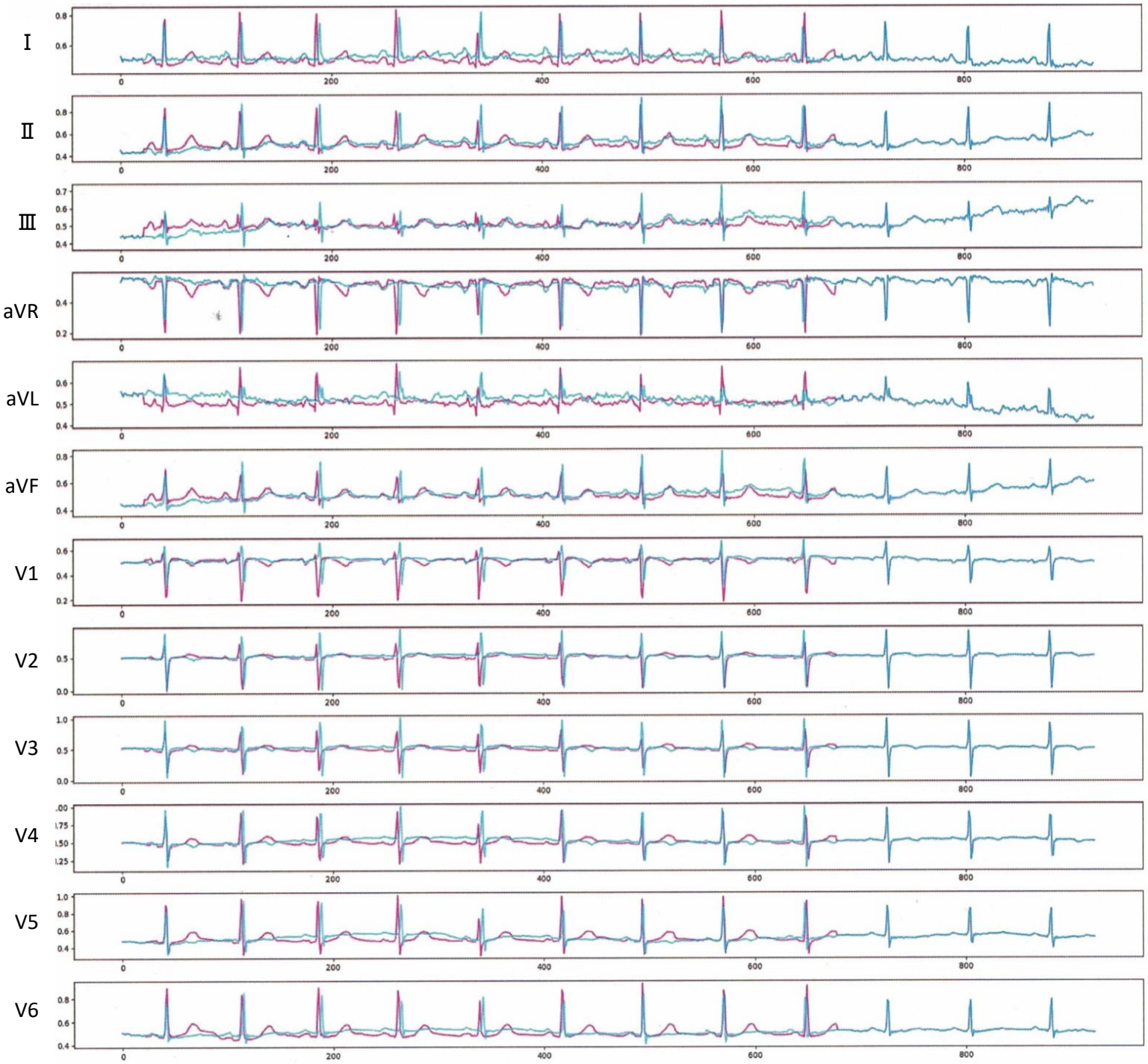


Native Guide Counterfactual – Expert A & B

Dataset: PTB-XL, Query: 394, NG: 627

Expert Prediction: Normal
AI Prediction: Normal (99,62%)

Expert Prediction: ST/T Change
Label: ST/T Change
AI Prediction: ST/T Change (94,82%)



Expert Counterfactual – Expert A

Expert Prediction: ST/T Change
Label: ST/T Change
AI Prediction: ST/T Change (94,82%)

Dataset: PTB-XL, Query: 394, NG: 627

Expert Counterfactual – Expert B

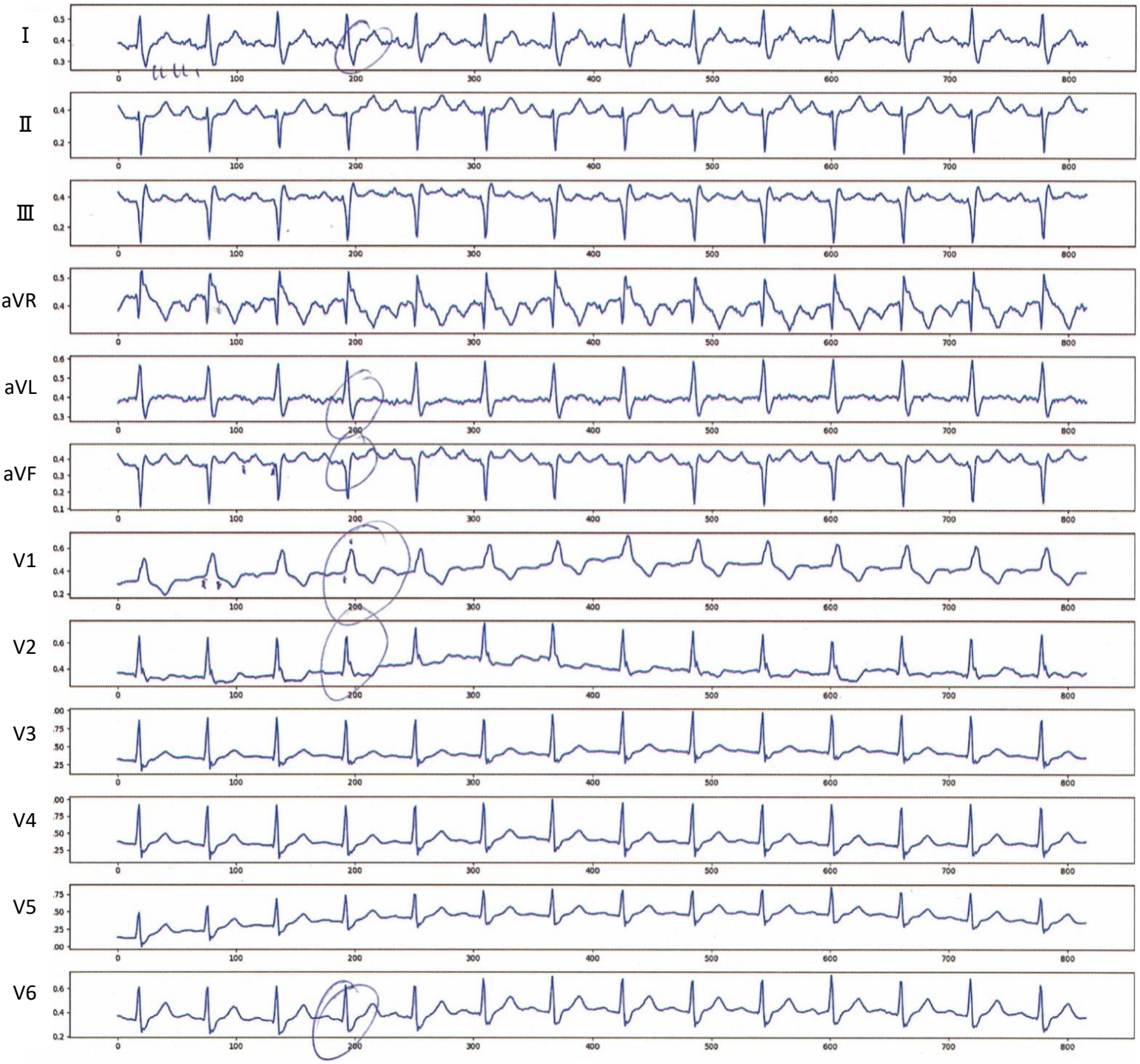
Expert Prediction: ST/T Change
Label: ST/T Change
AI Prediction: ST/T Change (94,82%)

Dataset: PTB-XL, Query: 394, NG: 627

Conspicuous Segments – Expert A

Expert Prediction: Conduction Disturbance
Label: Conduction Disturbance, Myocardial Infarction
AI Prediction: Conduction Disturbance, Myocardial Infarction (99,80%)

Dataset: PTB-XL, Query: 921, NG: 365



Conspicuous Segments – Expert B

Expert Prediction: Conduction Disturbance
Label: Conduction Disturbance, Myocardial Infarction
AI Prediction Conduction Disturbance, Myocardial Infarction (99,80%)

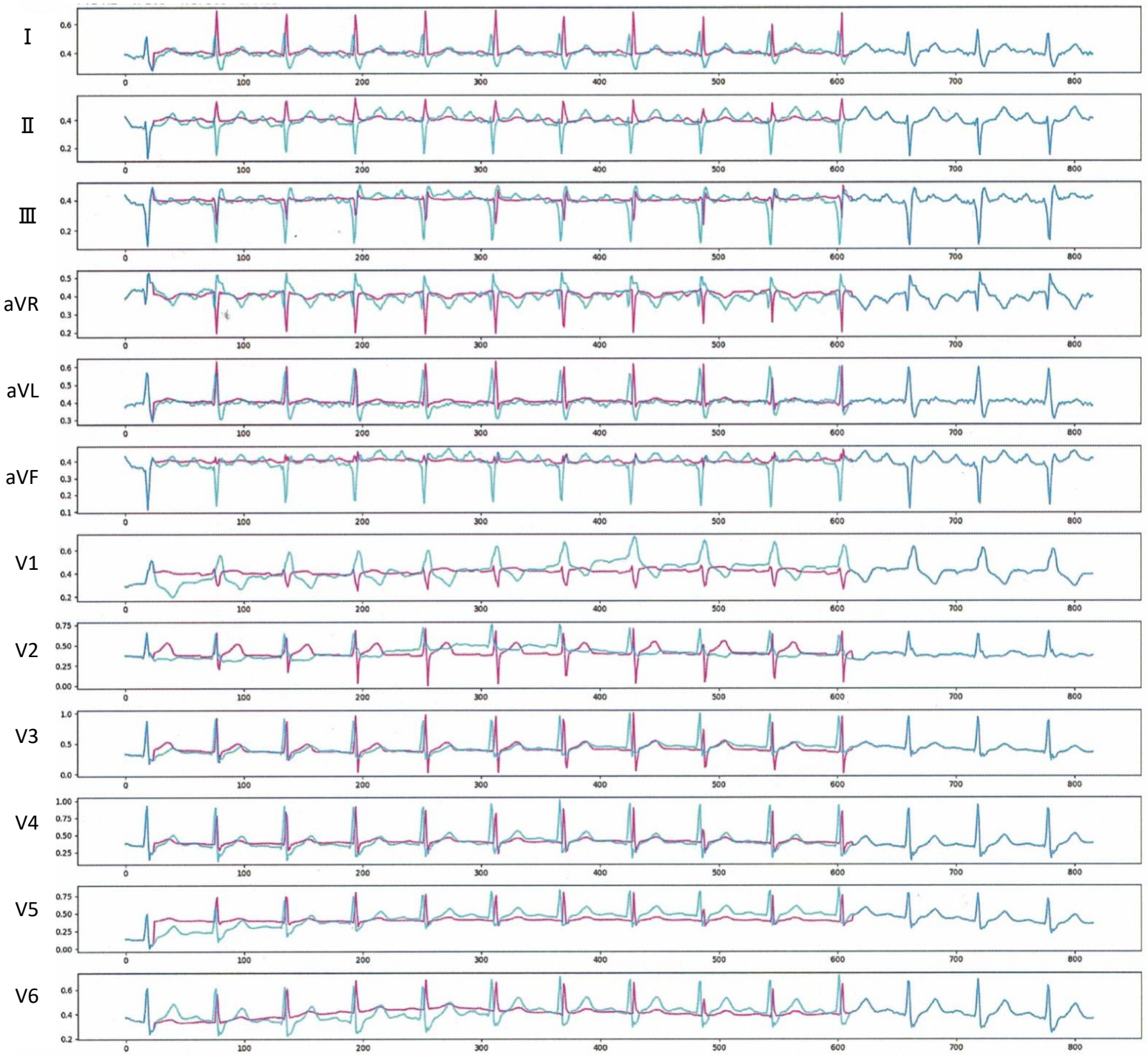
Dataset: PTB-XL, Query: 921, NG: 365

Native Guide Counterfactual – Expert A

Expert Prediction: Normal
AI Prediction: Normal (73,35%)

Expert Prediction: Hypertrophy or Normal
Label: Conduction Disturbance, Myocardial Infarction
AI Prediction: Conduction Disturbance, Myocardial Infarction (99,80%)

Dataset: PTB-XL, Query: 921, NG: 365

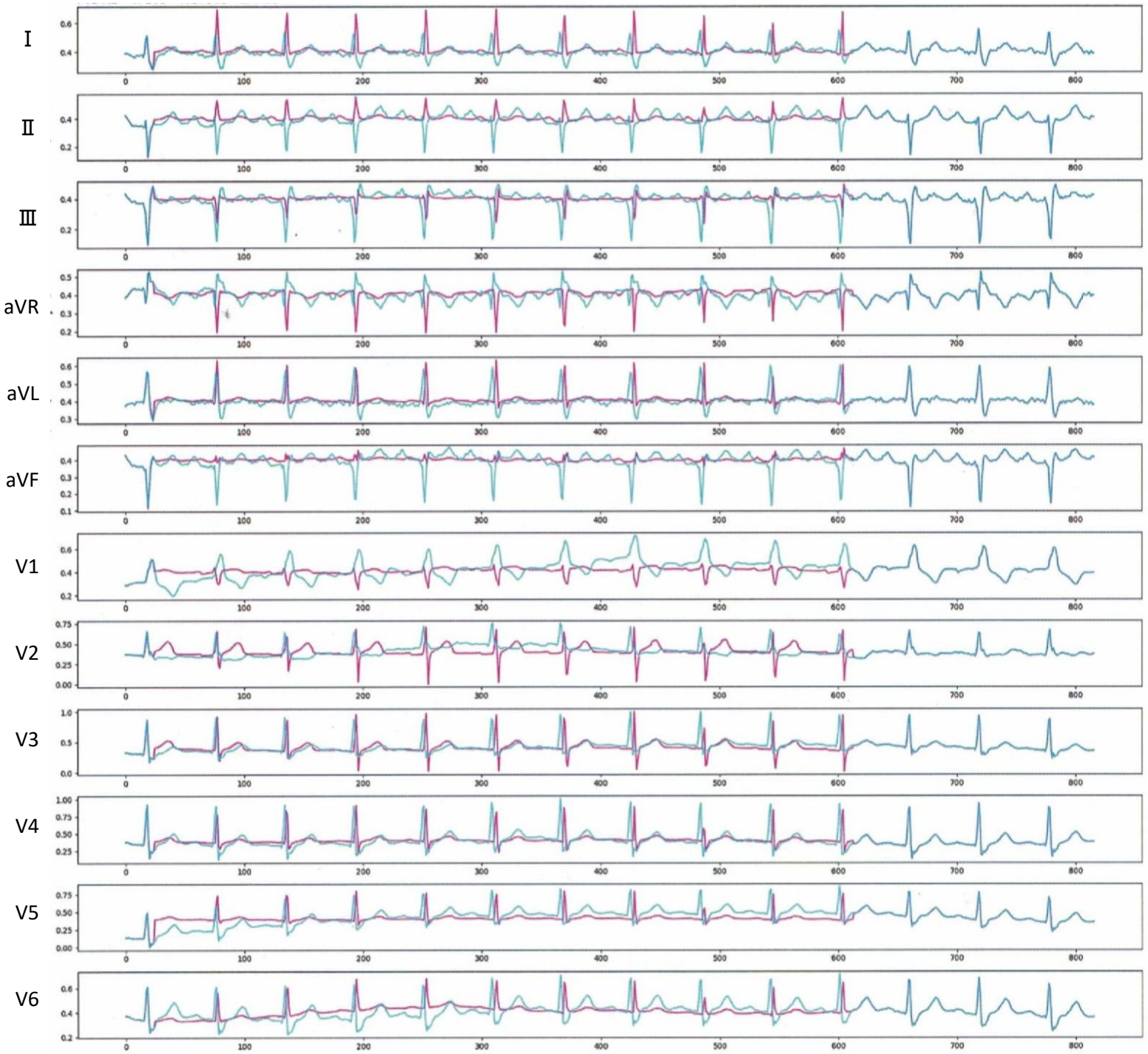


Native Guide Counterfactual – Expert B

Expert Prediction: Normal
AI Prediction: Normal (73,35%)

Expert Prediction: ST/T Change
Label: Conduction Disturbance, Myocardial Infarction
AI Prediction: Conduction Disturbance, Myocardial Infarction (99,80%)

Dataset: PTB-XL, Query: 921, NG: 365



Expert Counterfactual – Expert A

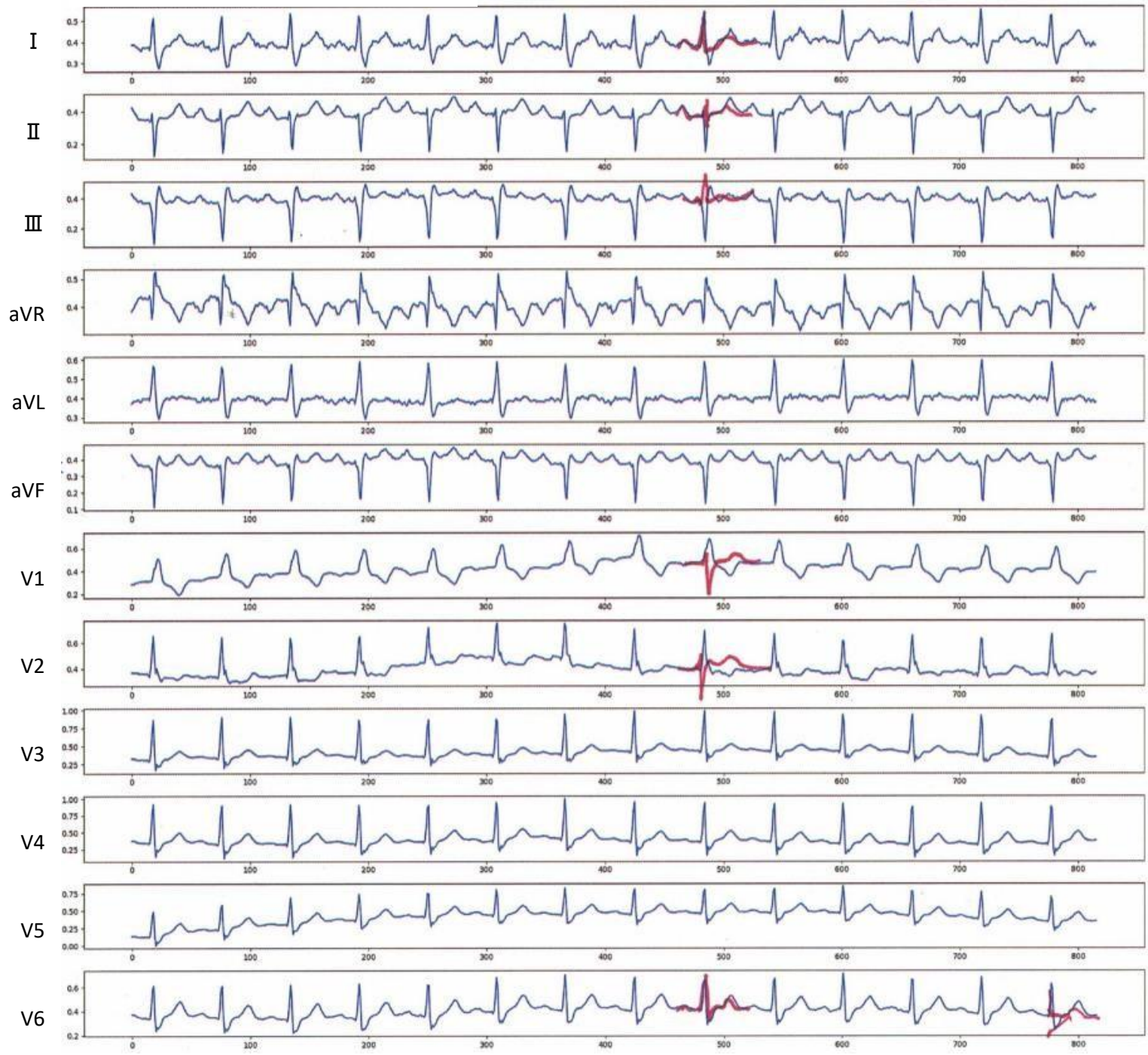
Expert Prediction: Conduction Disturbance
Label: Conduction Disturbance, Myocardial Infarction
AI Prediction: Conduction Disturbance, Myocardial Infarction (99,80%)

Dataset: PTB-XL, Query: 921, NG: 365

Expert Counterfactual – Expert B

Expert Prediction: Conduction Disturbance
Label: Conduction Disturbance, Myocardial Infarction
AI Prediction Conduction Disturbance, Myocardial Infarction (99,80%)

Dataset: PTB-XL, Query: 921, NG: 365



Conspicuous Segments – Expert A

Expert Prediction: Hypertrophy or Normal
Label: Normal
AI Prediction: Normal (99,71%)

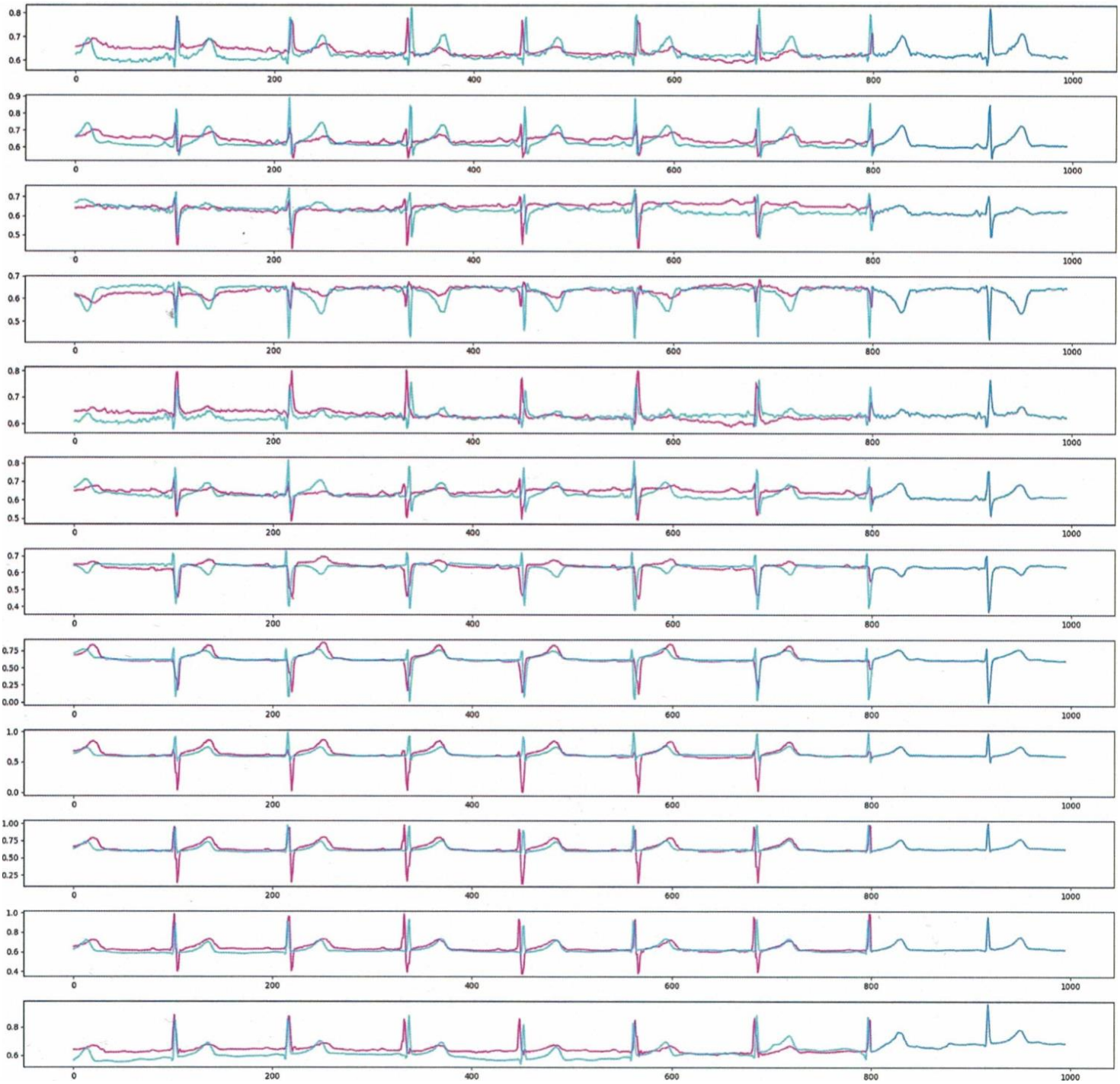
Dataset: PTB-XL, Query: 1212, NG: 425

Native Guide Counterfactual – Expert A

Expert Prediction: Normal
AI Prediction: Conduction Disturbance, Myocardial Infarction (87,32%)

Expert Prediction: Hypertrophy or Normal
Label: Normal
AI Prediction: Normal (99,71%)

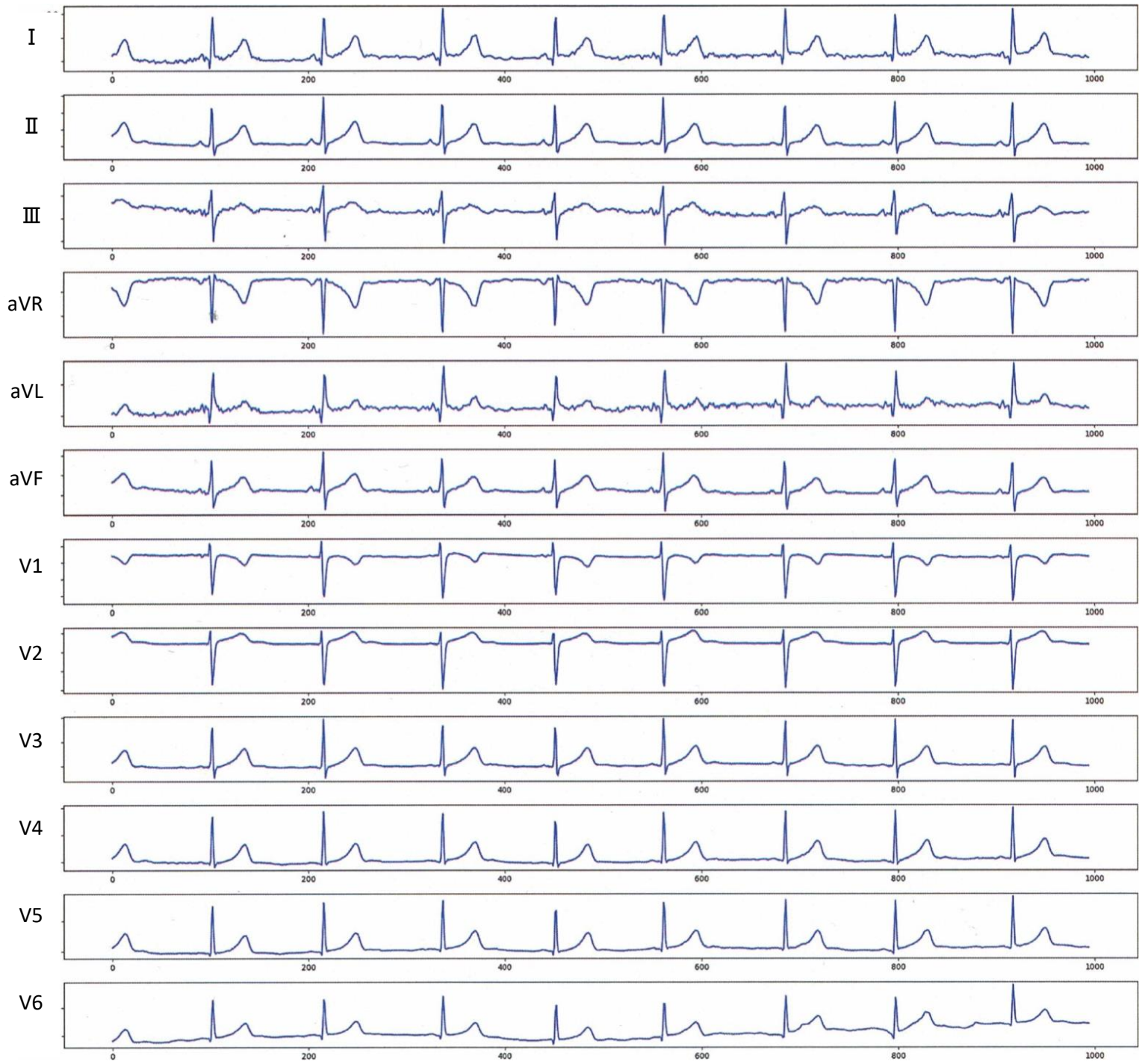
Dataset: PTB-XL, Query: 1212, NG: 425



No Expert Counterfactual – Expert A

Dataset: PTB-XL, Query: 1212, NG: 425

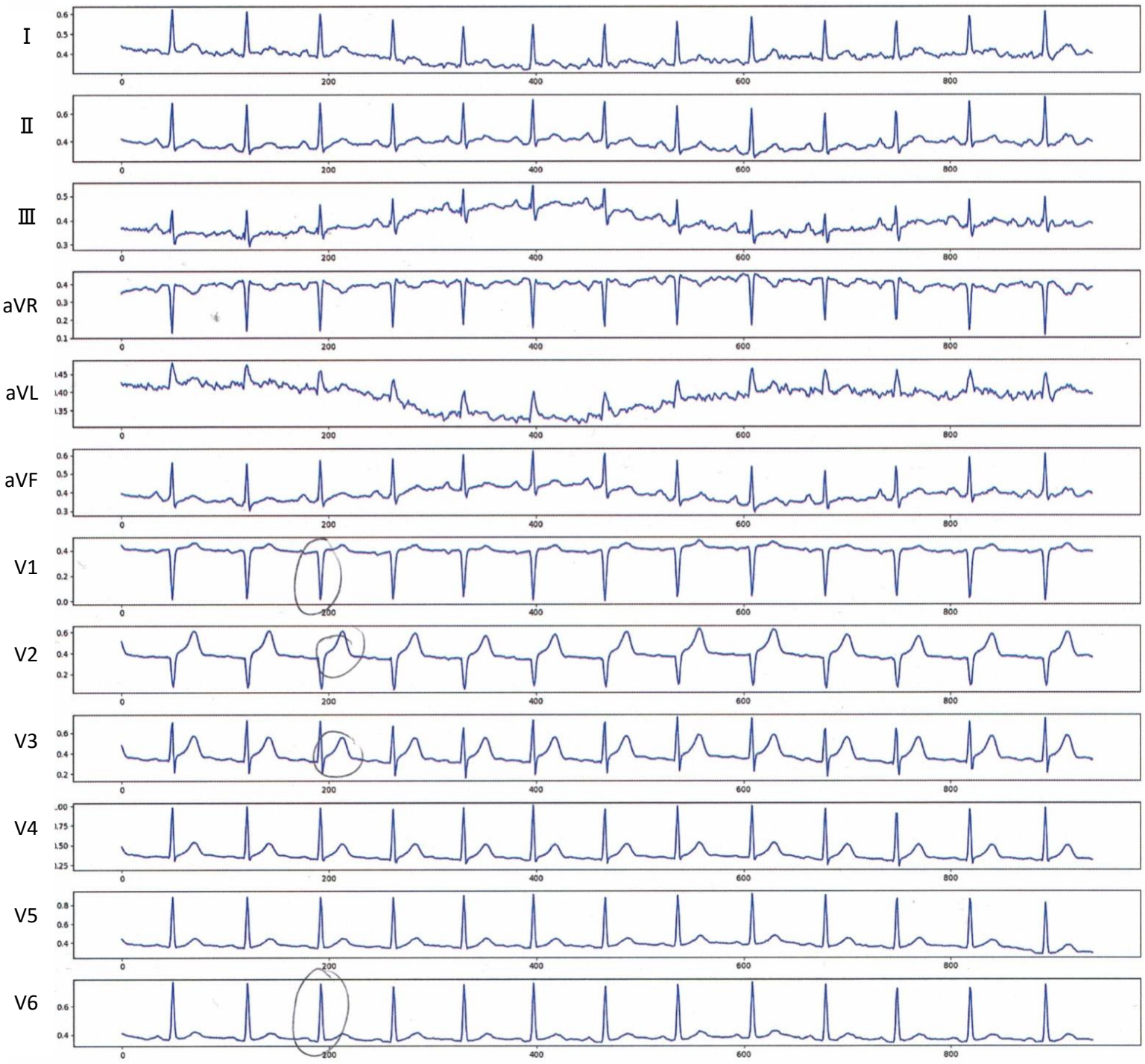
Expert Prediction: Hypertrophy or Normal
Label: Normal
AI Prediction: Normal (99,71%)



Conspicuous Segments – Expert A

Expert Prediction: Hypertrophy or Normal
Label: Myocardial Infarction
AI Prediction: Myocardial Infarction (77,21%)

Dataset: PTB-XL, Query: 474, NG: 221

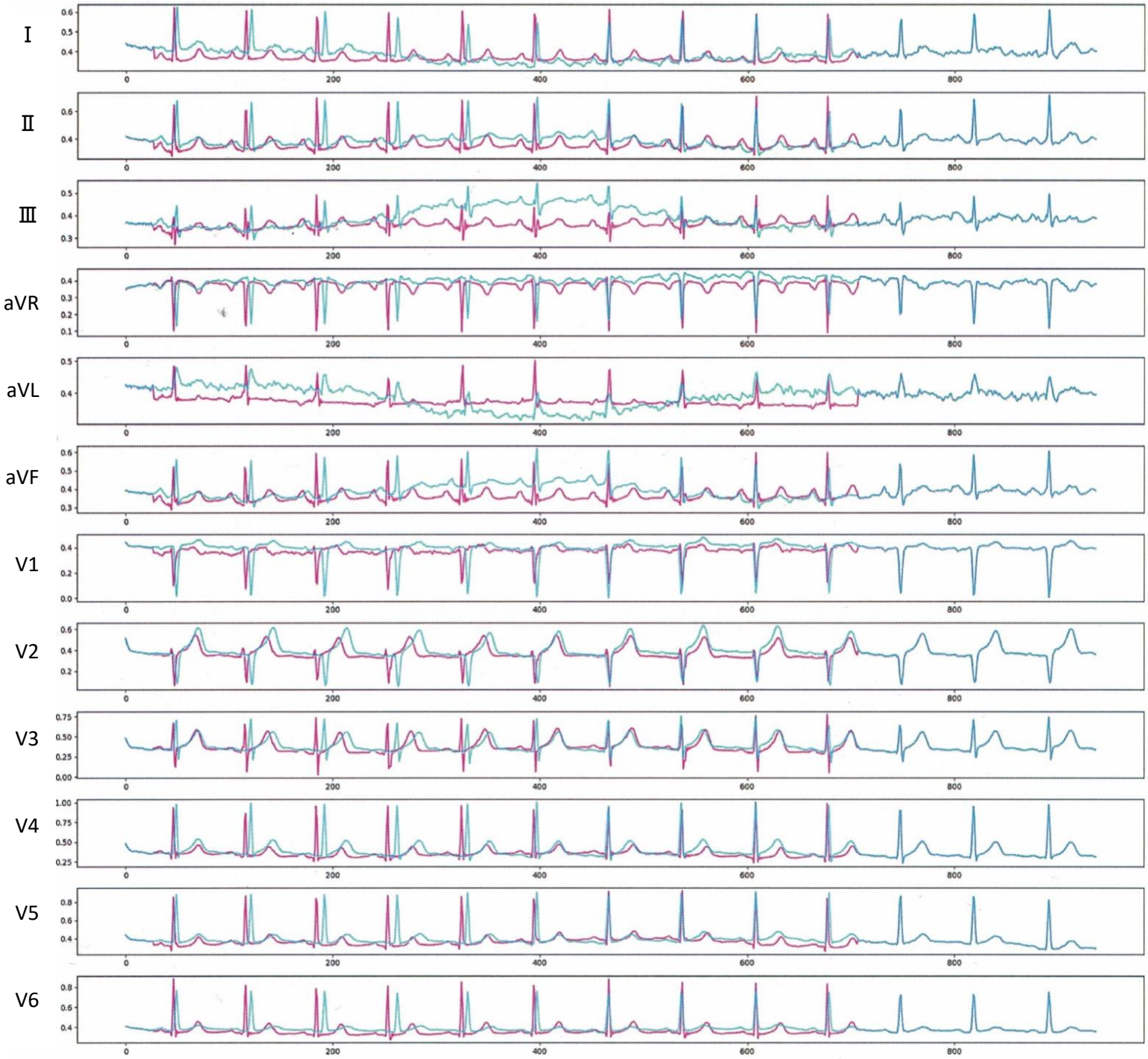


Native Guide Counterfactual – Expert A

Dataset: PTB-XL, Query: 474, NG: 221

Expert Prediction: Normal
AI Prediction: Normal (96,16%)

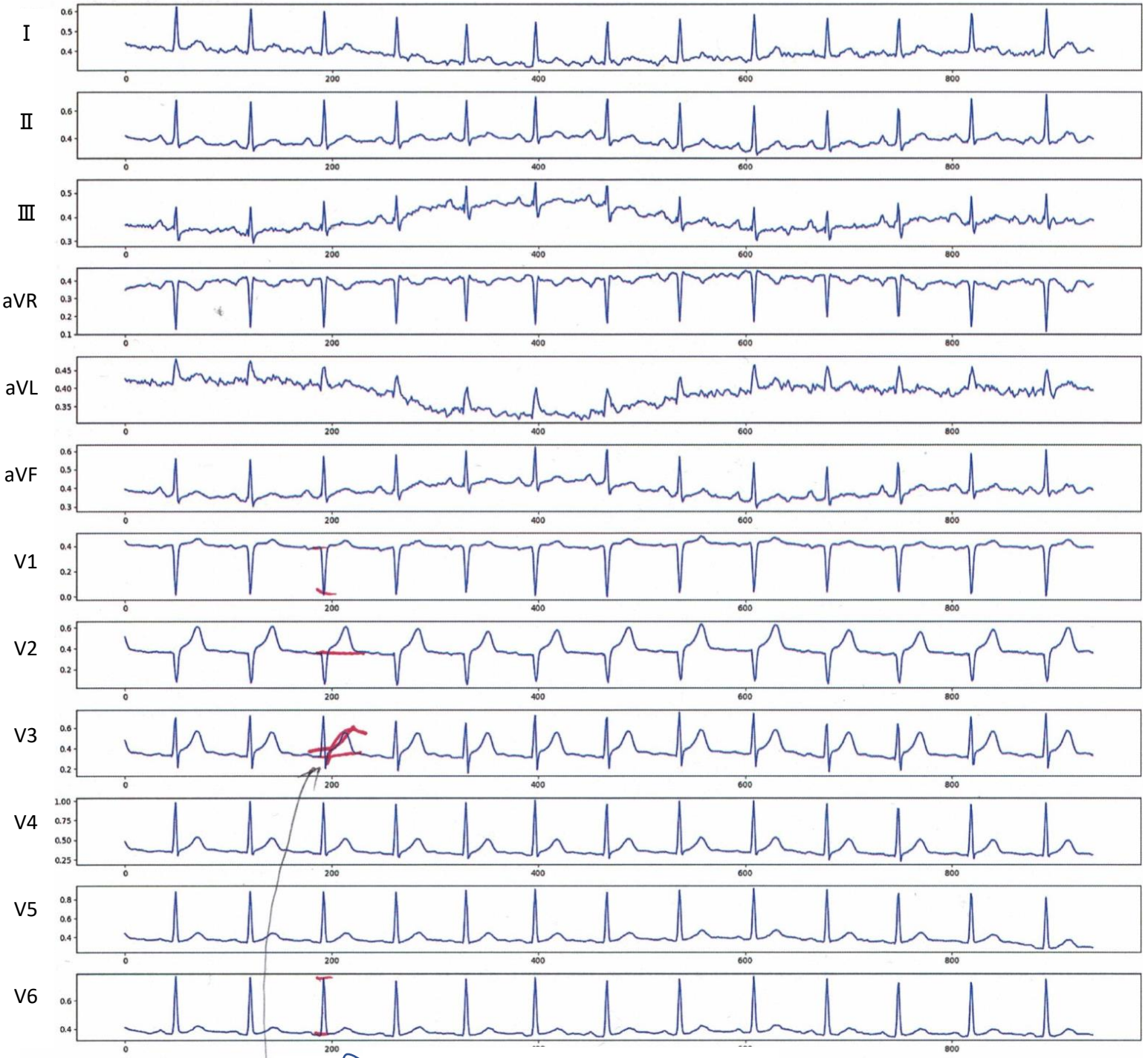
Expert Prediction: Hypertrophy or Normal
Label: Myocardial Infarction
AI Prediction: Myocardial Infarction (77,21%)



Expert Counterfactual – Expert A

Dataset: PTB-XL, Query: 474, NG: 221

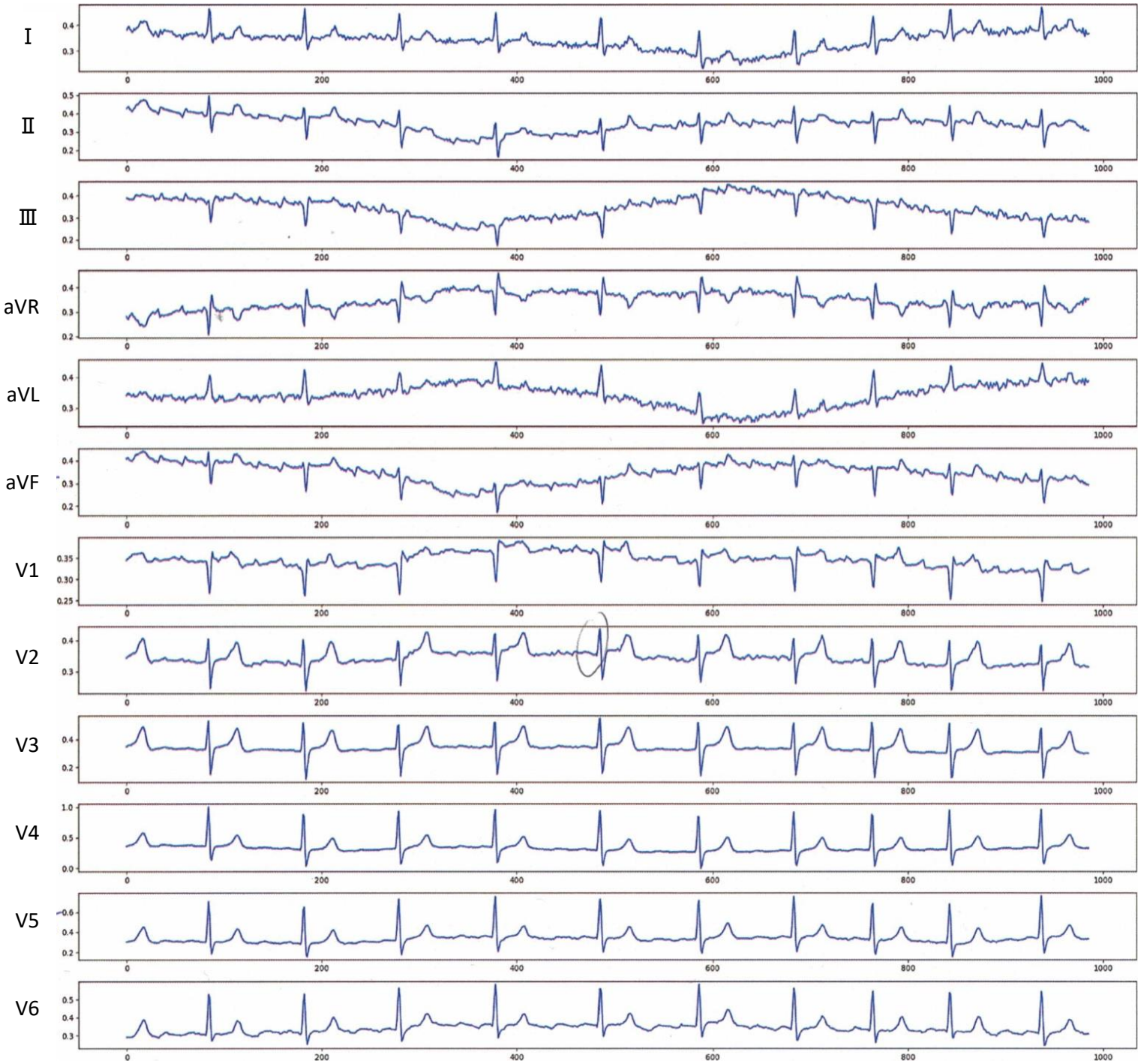
Expert Prediction: Hypertrophy or Normal
Label: Myocardial Infarction
AI Prediction: Myocardial Infarction (77,21%)



Conspicuous Segments – Expert B

Expert Prediction: Conduction Disturbance
Label: Conduction Disturbance
AI Prediction: Conduction Disturbance (97,12%)

Dataset: PTB-XL, Query: 1751, NG: 6

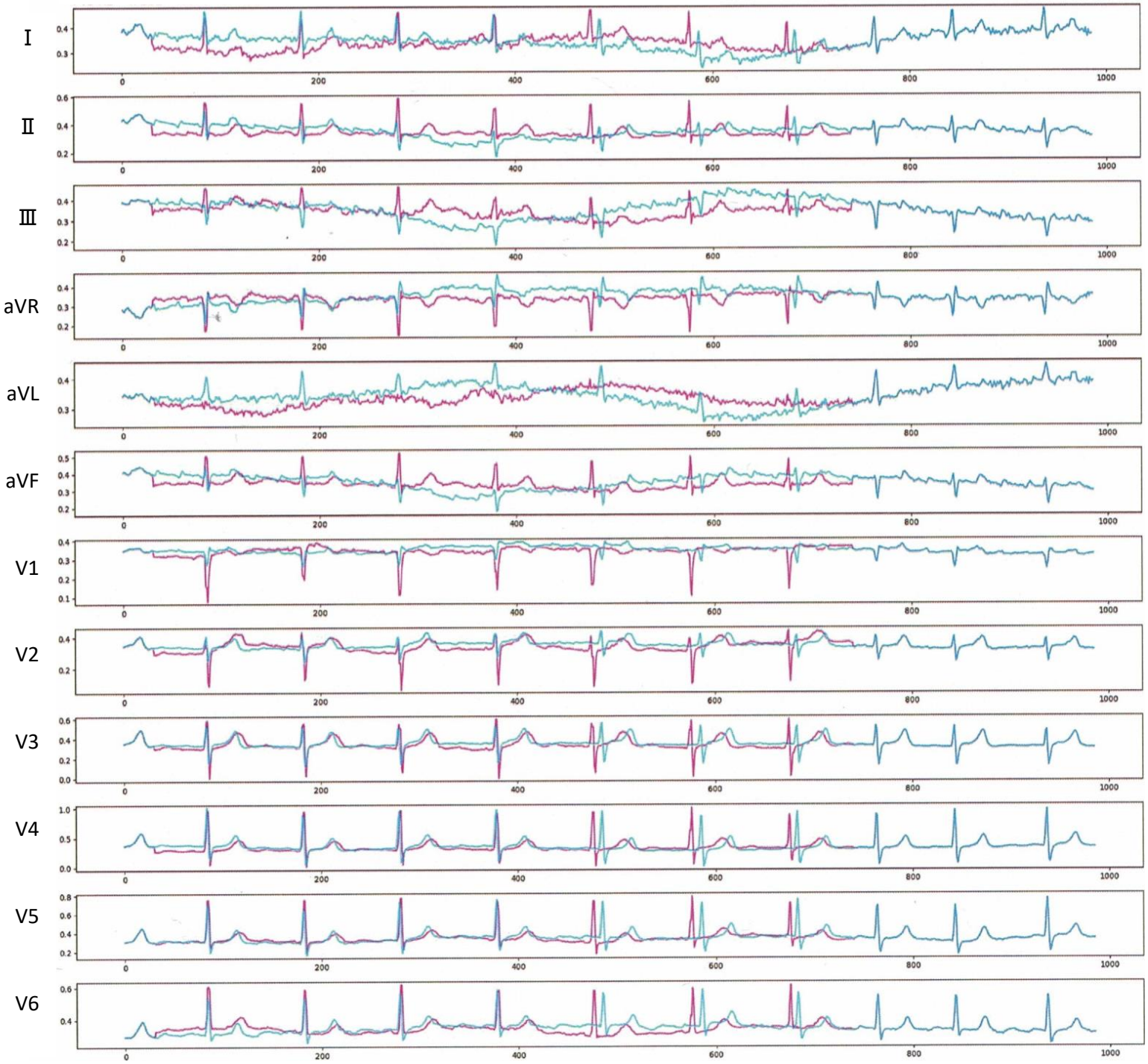


Native Guide Counterfactual – Expert B

Expert Prediction: Normal
AI Prediction Normal (88,44%)

Expert Prediction: Conduction Disturbance
Label: Conduction Disturbance
AI Prediction Conduction Disturbance (97,12%)

Dataset: PTB-XL, Query: 1751, NG: 6



Expert Counterfactual – Expert B

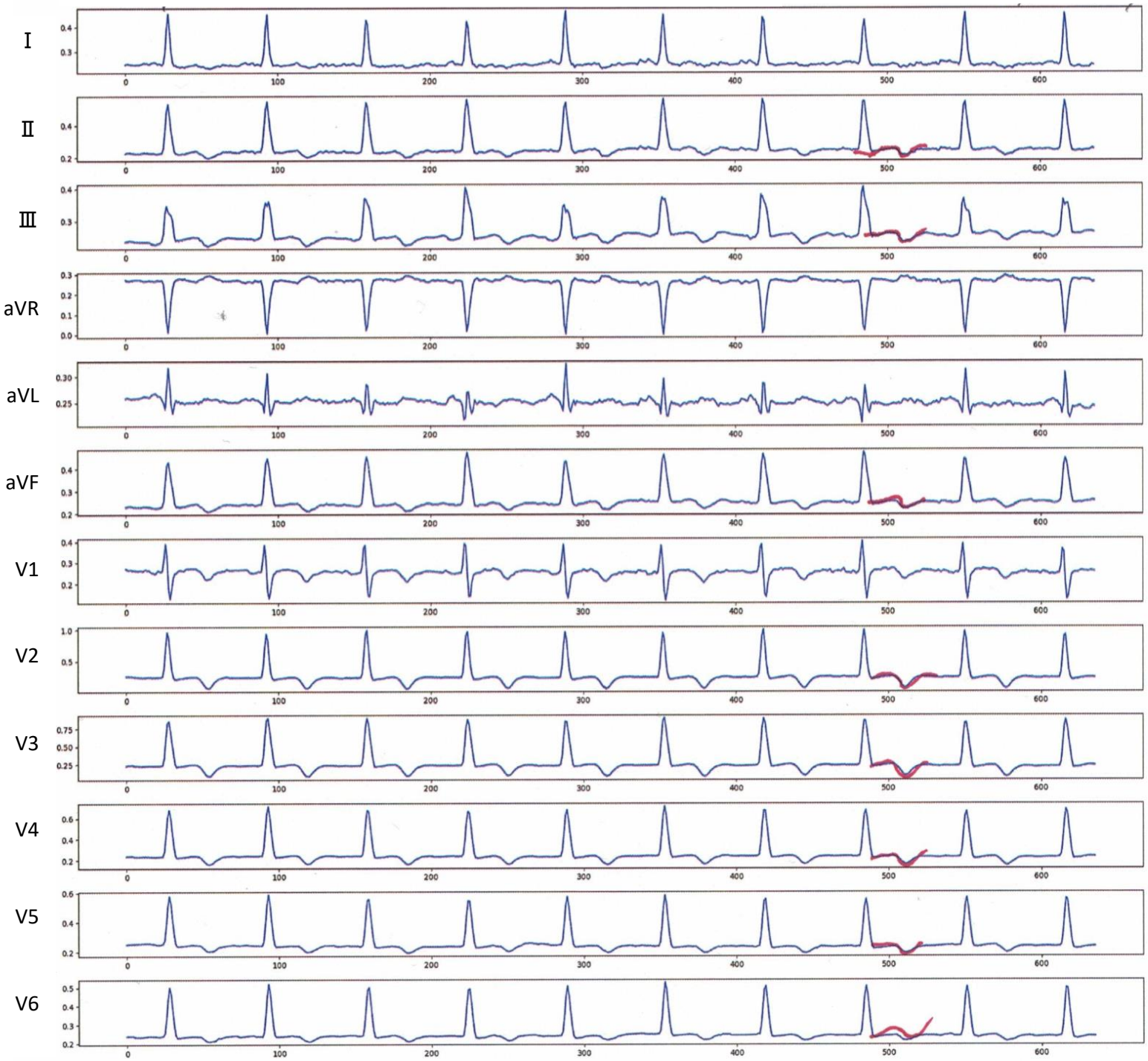
Expert Prediction: Conduction Disturbance
Label: Conduction Disturbance
AI Prediction Conduction Disturbance (97,12%)

Dataset: PTB-XL, Query: 1751, NG: 6

Conspicuous Segments – Expert B

Expert Prediction: ST/T Change
Label: Hypertrophy
AI Prediction: Conduction Disturbance (97,48%)

Dataset: PTB-XL, Query: 304, NG: 641

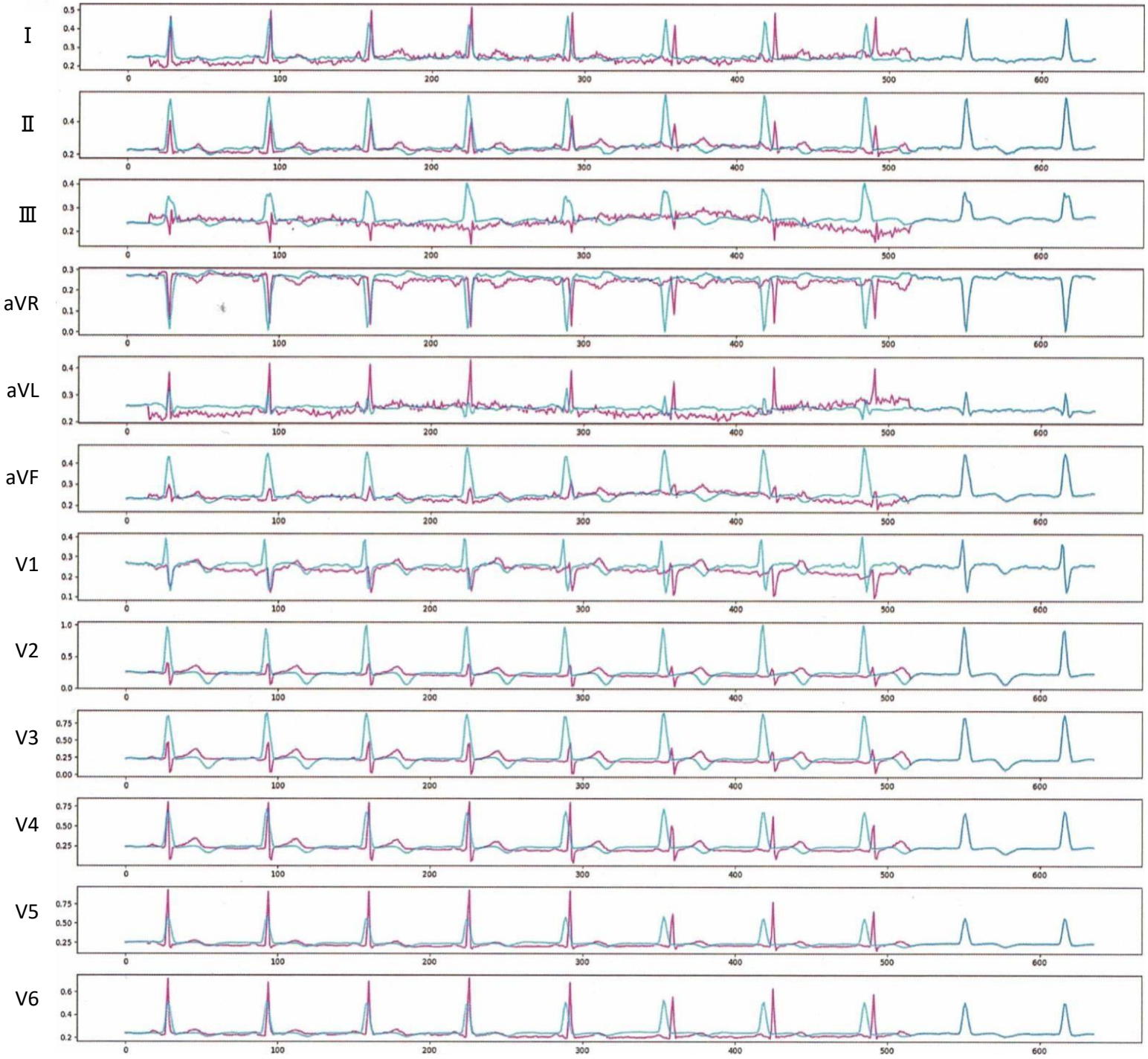


Native Guide Counterfactual – Expert B

Dataset: PTB-XL, Query: 304, NG: 641

Expert Prediction: Normal
AI Prediction: Normal (73,79%)

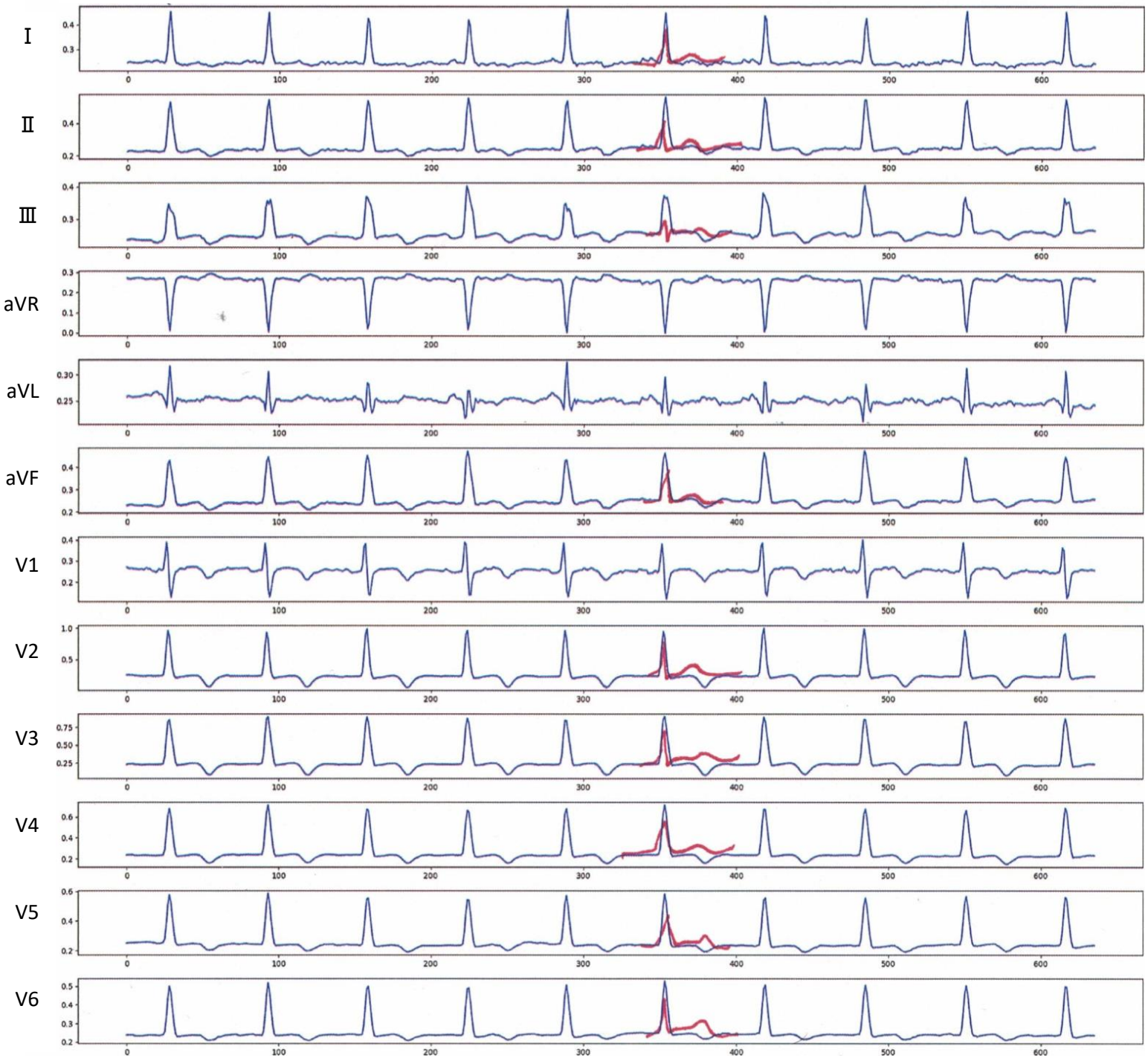
Expert Prediction: ST/T Change
Label: Hypertrophy
AI Prediction: Conduction Disturbance (99,71%)



Expert Counterfactual – Expert B

Expert Prediction: ST/T Change
Label: Hypertrophy
AI Prediction: Conduction Disturbance (99,71%)

Dataset: PTB-XL, Query: 304, NG: 641



B. Synchronization of ECG Plots

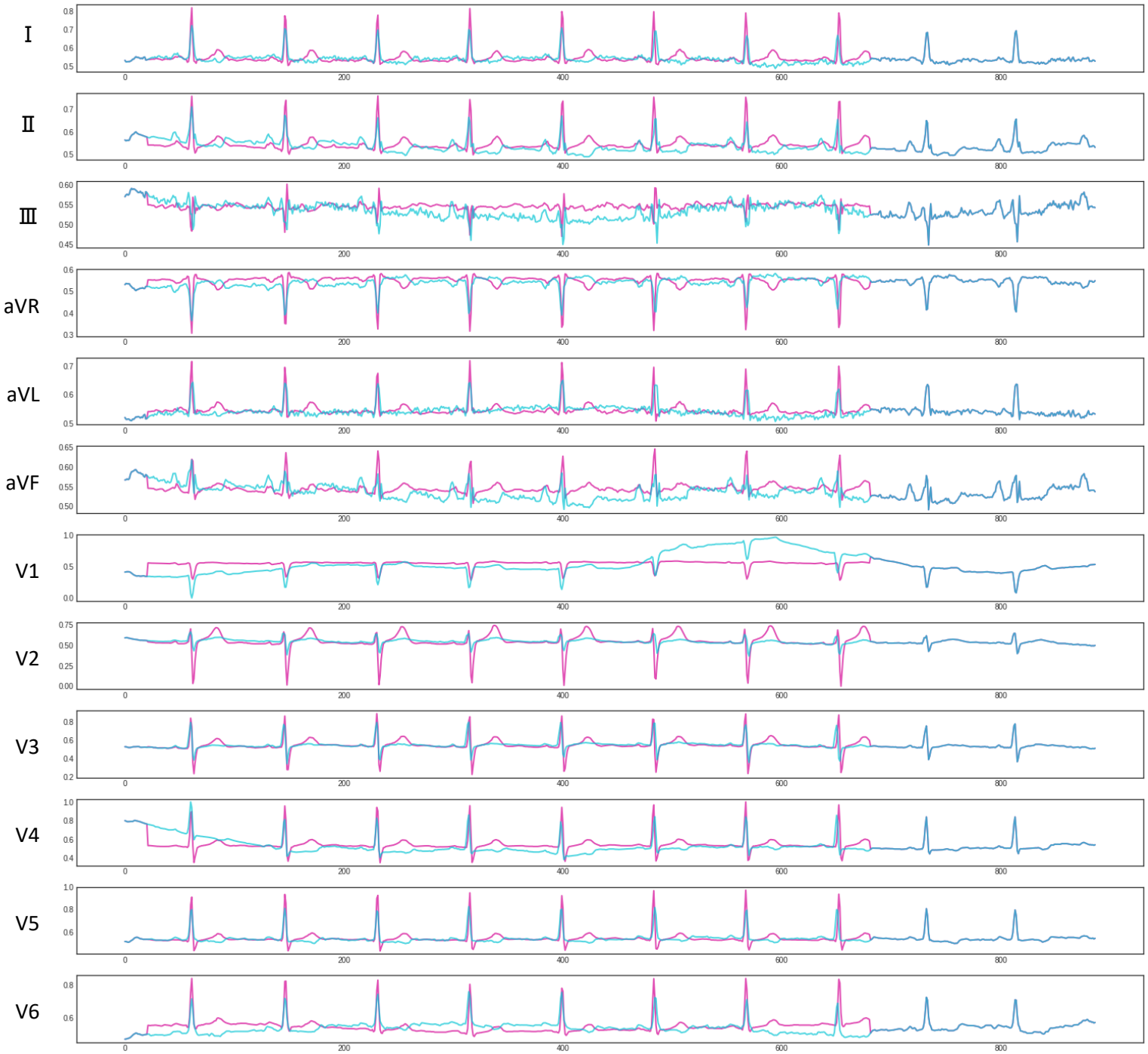
Synchronized Counterfactual

Counterfactual AI Prediction: Normal (95,28%)

Query Label: ST/T Change

Query AI Prediction: ST/T Change (95,40%)

Dataset: PTB-XL, Query: 930, NG: 935



C. Algorithms

Algorithm 4 Normalization norm(T)

1: $tmin \leftarrow getMin(T)$
2: $tmax \leftarrow getMax(T) - tmin$
3: **for each** t in T **do**
4: $\quad t \leftarrow t - tmin$
5: $\quad t \leftarrow div(t, tmax)$
6: **end for**
7: **return** T

div(x,y) divides x by y.
getMin() and getMax() get the minimal and maximal value of array input.

Algorithm 5 Wavelength Synchronization waveSync(T1, T2)

1: $wavelenT1 \leftarrow getAvgWavelen(T1)$
2: $wavelenT2 \leftarrow getAvgWavelen(T2)$
3: $s \leftarrow div(wavelenT2, wavelenT1)$
4: $T2sync \leftarrow \{\}$
5: **for** $k \leftarrow 1$ to $len(T2)$ **do**
6: $\quad T2sync.append(T2[int(k * s)])$
7: **end for**
8: $newLen \leftarrow len(T2sync)$
9: $del \leftarrow \{\}$
10: **for** $k \leftarrow 1$ to $newLen$ **do**
11: $\quad del.append(-(k + 1))$
12: **end for**
13: $T1 = delete(T1, del)$
14: **return** $T1, T2sync$

X[i] returns the data point of an array X at index i.
div(x,y) divides x by y.
delete(X,i) deletes elements from X by indices in i.
len() returns the length of array input.
getAvgWavelen() returns the average wavelength (from peak to peak) of an input ECG.

Algorithm 6 Peak Alignment shift(T1, T2)

1: $shiftT1 \leftarrow \{\}$
2: $shiftT2 \leftarrow \{\}$
3: $f1 \leftarrow fft(T1)$
4: $f2 \leftarrow fft(flip(T2))$
5: $crossCorr \leftarrow ifft(f1 * f2)$
6: $crossCorr \leftarrow real(crossCorr)$
7: $fshift \leftarrow fftShift(crossCorr)$
8: $center \leftarrow intdiv(len(T1), 2) - 1$
9: $shift \leftarrow center - argmax(fshift)$
10: $del = \{\}$
11: $delNeg = \{\}$
12: **for** $k \leftarrow 1$ to $len(T1)$ **do**
13: $\quad del.append(k + 1)$
14: $\quad delNeg.append(-(k + 1))$
15: **end for**
16: **if** $shift < 0$ **then**
17: $\quad T1 \leftarrow delete(T1, del)$
18: $\quad T2 \leftarrow delete(T2, delNeg)$
19: **else if** $shift > 0$ **then**
20: $\quad T1 \leftarrow delete(T1, delNeg)$
21: $\quad T2 \leftarrow delete(T2, del)$
22: **else**
23: $\quad T1 \leftarrow T1$
24: $\quad T2 \leftarrow T2$
25: **end if**
26: **return** $T1, T2$

flip(X) reverse the order of elements of X along axis 0.
len(X) returns the length of array input X.
delete(X,i) deletes elements from X by indices in i.
fft(X) und ifft(X) compute the one-dimensional discrete and inverse fourier transformation.
fftShift(X) shifts the zero-frequency component to the center of the input X.
intdiv(x,y) divides x by y and returns the integer number.
real(X) return the real part of the complex argument of input X.

Algorithm 7 Generation of $T^{'}$ by Point-for-Point Perturbation

1: $l_s \leftarrow 1$
2: $T^{'} \leftarrow T_q$
3: $c \leftarrow class(T_q)$
4: $c^{'} \leftarrow class(T_{NUN})$
5: $weightList \leftarrow getWeightList(T_{NUN})$
6: $maxWeightIndex \leftarrow argmax(weightList)$
7: $weightList[maxWeightIndex] \leftarrow -1$
8: $T'[maxWeightIndex] \leftarrow T_{NUN}[maxWeightIndex]$
9: **while** $class(T^{'}) = c$ **do**
10: $\quad l_s \leftarrow l_s + 1$
11: $\quad maxWeightIndex \leftarrow argmax(weightList)$
12: $\quad weightList[maxWeightIndex] \leftarrow -1$
13: $\quad T'[maxWeightIndex] \leftarrow T_{NUN}[maxWeightIndex]$
14: **end while**
15: **return** T'

class() returns the class of an input.
getWeightList() returns the weight of every data point of the input.
argmax() returns the index of the maximal number in an input array.
X[i] returns the data point of an array X at index i.