Explainable Artificial Intelligence: Transitioning DL Model Decisions to User-Understandable Features in Healthcare

Pavlo Radiuk[1], Oleksander Barmak[1], Eduard Manziuk[1], and Iurii Krak[2,3]

[1] *Khmelnytskyi National University, 11, Institutes str., Khmelnytskyi, 29016, Ukraine*
[2] *Taras Shevchenko National University of Kyiv, 64/13, Volodymyrska str., Kyiv, 01601, Ukraine*
[3] *Cybernetics Institute, 40, Glushkov ave., Kyiv, 03187, Ukraine*

Abstract

Modern artificial intelligence (AI) solutions often face problems of the "black box" nature of deep learning (DL) models, which limits their transparency and trustworthiness in critical medical applications. In this study, we propose and evaluate a scalable approach for enhancing the interpretability of DL models in medical signal and image processing by translating complex model decisions into features that are understandable to healthcare professionals. The proposed approach was tested on two medical datasets: ECG signals for arrhythmia detection and MRI scans for heart disease classification. The performance of the DL models was compared with expert annotations, using Cohen's Kappa coefficient as the primary metric to assess agreement. The study found strong agreement between DL model predictions and expert annotations, with Cohen's Kappa coefficients of 0.89 for the ECG dataset and 0.80 for the MRI dataset, demonstrating the usefulness of the approach in providing reliable and interpretable results. In sum, the proposed scalable approach significantly improves the interpretability of DL models in medical applications.

Keywords

Explainable artificial intelligence, deep learning, medical signal processing, medical image analysis, model interpretability, Cohen's Kappa

1. Introduction

The rapid development of AI has made it important to explain the decisions made by AI, or in other words, to develop understandable Machine Learning (ML). The term "explainable artificial intelligence" (XAI) [1] suggests an AI system, which decisions are understandable to the average user. This term contrasts with the "black box" concept [2], where the decisions generated by AI are incomprehensible to the end user. It is also worth noting that XAI implements the «right to explanation» [3], that is, the right to have a clear explanation of the result of the algorithm's work. This right applies to each of us when the algorithm's decision directly affects a person. Such rights are already being developed, although the general "right to explanation" is still under discussion. In the information society, the "right to explanation" is becoming an extremely important concept, as digital technologies, AI, and ML will continue to be actively applied to solving various problems of human activity [4, 5].

In general, it is believed that XAI adheres to three principles: transparency, interpretation, and explanation [6]. We can talk about the inherent transparency of XAI if the developer can describe and explain how the model forms and updates parameters from statistical training data and how it makes predictions on new data [7]. By interpretation of XAI, we mean understanding how the AI model forms its output data and explaining its decisions to people [8]. Explanation in XAI is an

ProfIT AI 2024: 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024), September 25–27, 2024, Cambridge, MA, USA

radiukp@khmnu.edu.ua (P. Radiuk); barmako@khmnu.edu.ua (O. Barmak); manziuk.e@khmnu.edu.ua (E. Manziuk); iurii.krak@knu.ua (Iu. Krak)

0000-0003-3609-112X (P. Radiuk); 0000-0003-0739-9678 (O. Barmak); 0000-0002-7310-2126 (E. Manziuk); 0000-0002-8043-0785 (Iu. Krak)

important concept but without a clear definition. It is believed that AI explanation in a broad sense is a set of features that influenced the decision (i.e., classification or prediction) for a specific case [9]. If AI-based approaches meet these requirements, then they are said to provide the basis for justifying decisions, tracking and verifying them, and improving and researching new facts [10].

XAI issues are especially critical in areas such as medicine, defense, finance, and law, where it is important to understand AI decisions and trust them [11]. Today there are many approaches that provide decent results in various tasks of such sensitive areas of human activity [12]. In general, DL methods provide better results compared to traditional ML methods for solving problems with heterogeneous data. In particular, convolutional neural network (CNN) models [13] are state-of-the-art for computer vision tasks [14], and transformer models are state-of-the-art for natural language processing tasks [15]. However, as already mentioned, decisions made by DL methods are not always transparent and understandable.

In this study, we address the problem of explaining decisions made by AI. In our previous work [16], we proposed an approach to explaining the results of a DL model based on visual analytics with a transition matrix from the DL model to the ML model. However, the use of this approach [16] to explain DL decisions assumes the presence of two corresponding models, DL and ML. However, in practice, there are often cases when there is a DL model, but there is no corresponding ML model. In this case, we cannot apply the specified approach to explaining the decisions made by the DL model. In this study, we propose an extension of our previous approach [16] to consider the absence of an ML model corresponding to the DL model.

Therefore, this work aims to apply our previous approach to the case of medical data processing, where there is no ML model like the DL model; instead, we consider a set of features that are understandable to a healthcare expert. Using these features, we propose to interpret the decisions obtained by the DL model. Finally, the main contribution of this work is the scalable approach to explain the results obtained by DL models, based on features understandable to a healthcare expert (end-user) for medical signal and image processing tasks.

The structure of the article is as follows. Section 2 provides an analysis of sources on explainable artificial intelligence for critically important tasks. Section 3 describes the proposed approach to presenting decisions made by DL models using features understandable to the physician, given the solution of medical signal and image processing problems. Section 4 presents the results of computational experiments and their interpretation.

2. Related works

The field of XAI is experiencing significant advancements, particularly in the development of methods to enhance the transparency of AI models in the healthcare domain. Researchers are actively exploring various approaches, including the construction of feature models and the use of manually crafted features to provide clearer explanations of AI decisions. These efforts are crucial for fostering trust and reliability in AI-driven clinical applications, as they bridge the gap between complex model outputs and the need for interpretable, actionable insights in healthcare.

As an example, Bassiouny et al. [17] present an innovative approach to diagnosing neonatal lung diseases by training an object detection model, faster-RCNN, to identify seven key lung ultrasound features rather than making direct diagnostic predictions. This methodology enhances the interpretability of the results and keeps clinicians in control by providing annotated images to support their diagnostic decisions. The study demonstrates that the model surpasses single-stage detectors like RetinaNet, achieving high mean average precision, thus balancing performance with trustworthiness in medical practice.

In their review, Salahuddin et al. [18] explore various interpretability methods for deep neural networks in medical image analysis, emphasizing that these methods aim to enhance transparency and trust in AI systems. They highlight that while these interpretability techniques provide valuable insights, they are often approximations and may not fully capture the true decision-making processes of the models, necessitating cautious application in clinical settings. In addition, Chan et

al. [19] developed and compared three machine learning models to predict long-term mortality in critically ill ventilated patients, finding that boosting algorithms, random forest, and logistic regression achieved similar performance.

Similarly, Lu et al. [20] propose a comprehensive workflow that includes a step where medical professionals label differential diagnosis features according to medical guidelines, effectively blacklisting irrelevant features extracted from electronic health records. This approach aims to «reduce workloads of clinicians in human-in-loop data mining» by focusing on feature oversight rather than full prediction, thus enhancing the trustworthiness and efficiency of the AI model.

In [21], Moreno-Sánchez et al. enhances a heart failure survival prediction model by integrating explainable AI techniques, aiming to balance predictive performance and interpretability. This approach provides transparency by explaining feature contributions to predictions, making the model's decision-making process clearer for clinicians. Consequently, it fosters greater trust and practical adoption in clinical settings.

Pintelas et al. [22] introduce a novel framework for 3D image recognition that utilizes interpretable features such as lines, vertices, and contours to enhance explainability. This approach is particularly promising for medical imaging, achieving performance comparable to state-of-the-art black-box models while maintaining transparency. However, the development of interpretable methodologies for 3D image segmentation remains an emerging area of research, with most existing techniques originally designed for 2D image classification tasks.

Consequently, the analysis of related sources has revealed the absence of clear methodologies for constructing feature models. Consequently, the primary objective of this study is to elucidate the decision-making processes of DL models in addressing the challenges of processing medical signals and images. To fulfill this aim, it is essential to develop a new scalable visual analytics approach, utilizing a transition matrix to bridge the DL model's decision-making with the task of explaining these decisions in the context of medical signal and image processing.

3. Methods and materials

3.1. Basic approach

In our previous work [16], the problem of explaining decisions made by DL models is considered according to the idea illustrated in Figure 1.

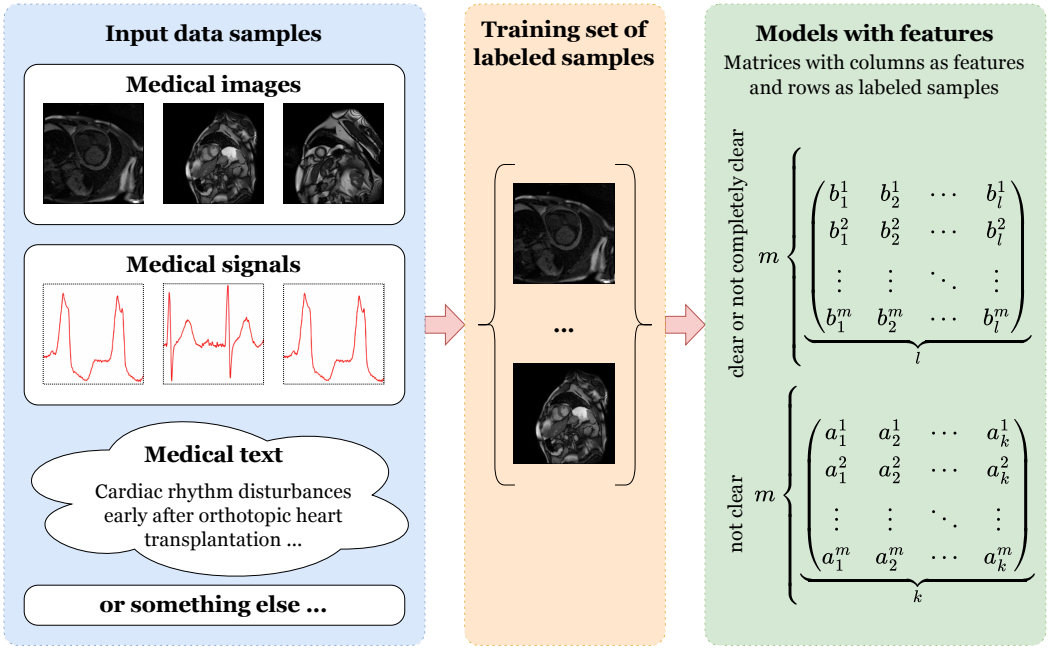


Figure 1: This diagram illustrates the process of converting various input medical data samples, such as images, signals, text, etc. into a training set of labeled samples, which are then used to generate

models with features categorized as either clear or unclear to the end user, forming the basis for explainable AI in medical applications.

Let us have input information that can be presented in the form of i) images; ii) signals; iii) text fragments, etc. (Figure 1). The end user, who is assumed to be a specialist in a specific subject area, empirically, according to their own experience and knowledge of the subject area, highlights certain areas of attention (receptive fields) on this input information and forms a so-called set of features that characterize the objects of the input information. According to the current appearance of these features, a decision is made about the state in which these areas of attention are located: qualitatively or quantitatively. The set of states of these features is intended to form a so-called transparent solution, for example, assigning to a certain class.

The process described above involves the formation of ML models, which have all the necessary features of understandable AI: transparency, interpretability, and explainability. Otherwise, these features (areas of attention) are formed according to certain algorithms (DL models) and, as a result, are not entirely clear, or not at all clear to the end user.

It is worth noting that there are also intermediate cases when the signs are «in the middle» between the indicated cases, such as:

- Decomposition of understandable areas of attention into «incomprehensible» signs, both with the possibility of reverse transformation and without such possibility.
- In addition to features understandable to the public of experts, there may be separate features (or combinations of previously obtained ones) that are understandable to more experienced experts or are based on intuition; here, for each specific case, the community decides whether these cases are transparent or not.

As a result of the above, let us have one training sample and two models with features. One model is built by the DL model, and the other model has features formed by an expert.

Next, let us formalize the problem under consideration. We represent the features of the DL model in the form of matrix A of dimension $m \times k$ and the features of the other model in the form of matrix B of dimension $m \times l$ as follows:

$$m\left\{\underbrace{\begin{pmatrix} a_1^1 & a_2^1 & \cdots & a_k^1 \\ a_1^2 & a_2^2 & \cdots & a_k^2 \\ \cdots & \cdots & \cdots & \cdots \\ a_1^m & a_2^m & \cdots & a_k^m \end{pmatrix}}_{k}\right. = A, \qquad (1) \qquad m\left\{\underbrace{\begin{pmatrix} b_1^1 & b_2^1 & \cdots & b_l^1 \\ b_1^2 & b_2^2 & \cdots & b_l^2 \\ \cdots & \cdots & \cdots & \cdots \\ b_1^m & b_2^m & \cdots & b_l^m \end{pmatrix}}_{l}\right. = B, \qquad (2)$$

where m is the number of vectors obtained from the training sample during DL model training, k represents the number of features, l stands for the number of features.

We emphasize once again that the features formalized in formulas (1)–(2) are obtained from the same training sample. We also note that in general k can be equal to, less than, or greater than l.

In practical problems that are modeled in this way, that is, in the presence of two mappings for the same objects for different sets of features, it is often necessary to express feature vectors of different dimensions through each other. In other words, consider the problem where for different matrices A and B it is necessary to find such a matrix T that the following equality holds:

$$B = TA, \qquad (3)$$

where T is the transition matrix between matrices A and B.

Note that in linear algebra, formula (3) is a usual change of basis of a vector space, and if the condition $m = k = l$ is met, finding matrix T is trivial, that is:

$$T = BA^{-1}. \qquad (4)$$

For the case under consideration, $m \neq k \neq l$, the inverse matrix does not exist, and therefore it is proposed to apply a generalization of the inverse matrix – the pseudo-inverse matrix. We propose to find such a matrix T of dimension $k \times l$, which will provide the transition between matrices A and B:

$$AT \approx B. \qquad (5)$$

Note that the approximation in formula (5) is established with respect to the Euclidean norm in the feature space of the matrices. It is proposed to find matrix T as follows:

$$T \approx A^{+}B. \tag{6}$$

In practice, it is proposed to define A^{+} using SVD decomposition [24], even though other methods are described in the previous work [16]:

$$A^{+} = V\Sigma^{+}U^{T}, \tag{7}$$

where $A = V\Sigma U^{T}$ is the singular value decomposition of matrix A, the matrix Σ^{+} is formed by transposing the matrix Σ and replacing all its non-zero values of diagonal elements with inverse ones:

$$\Sigma^{+} = [D^{+} : 0], \quad m \geq k;$$

$$\Sigma^{+} = \begin{bmatrix} D^{+} \\ \cdots \\ 0 \end{bmatrix}, \quad m < k;$$

$$D^{+} = diag(\sigma_1^{+}, \sigma_2^{+}, \sigma_3^{+}, \ldots, \sigma_t^{+}), \quad \sigma_i^{+} = \begin{cases} \dfrac{1}{\sigma_i^{+}}, & \sigma_i > 0; \\ 0, & \sigma_i = 0. \end{cases}$$

Therefore, for an arbitrary row vector of features a_j^{*}, $i = \overline{1,k}$, obtained by the model defined by matrix A, the corresponding row vector of features b_i^{*}, $j = \overline{1,l}$, by the model defined by matrix B, is determined using the obtained transition matrix T as follows:

$$b_i^{*} = a_j^{*}T, \quad i = \overline{1,k}, \quad j = \overline{1,l}. \tag{8}$$

The approach described above (1)–(8) somehow correlates with approximation, that is, the description by one function, even given in tabular form, of a given form of another function, perhaps also in tabular form.

There are several approaches to data approximation. One of them consists in approximating a complex function with a simpler function, which is used for all tabular values, but it is not necessary that it passes through all points. This approach is also called curve fitting, which is sought to be carried out so that its deviation from the tabular data is minimal. The authors propose to use the transition matrix T according to formula (4) between two feature models, presented in the form of matrices, for the same set of input data as such a function.

Figure 2 briefly shows the main steps of the basic approach, first proposed in our previous work [16], to obtaining the transition matrix T.

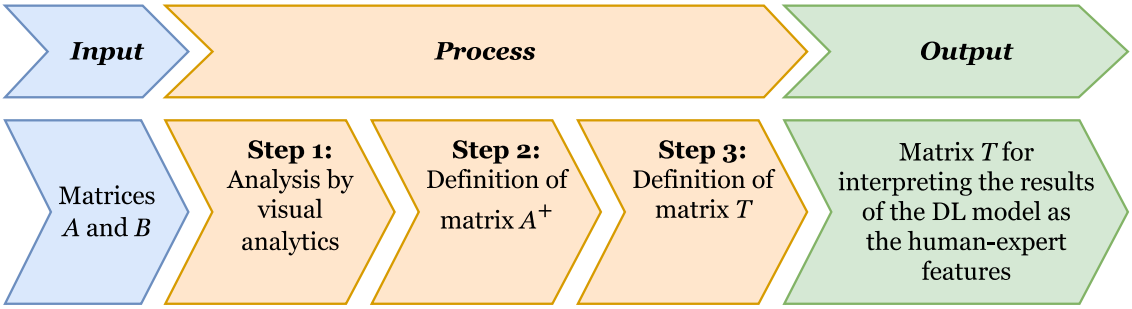


Figure 2: Diagram of the basic approach to obtaining the transition matrix T.

Below we will briefly describe the steps of the approach illustrated in Figure 2.

First, we extract two matrices, A from the DL model and B from the ML model, both representing the same data samples.

Step 1: Use a visual analytics tool to map these feature vectors as points on a plane, ensuring their relative positions match across models.

Step 2: Compute the pseudo-inverse matrix A^{+}.

Step 3: Calculate the transition matrix T.

Finally, use matrix T to translate DL results into features understandable by the ML model.

3.2. The proposed scalable approach

This paragraph describes a new way of constructing matrix B without an ML model, which will allow applying the approach from Figure 2 when no ML model corresponds to the DL model.

The proposed scalable approach is aimed at simplifying complex, hard-to-understand features from a DL model into a more user-friendly form, making the results easier to interpret. The extracted feature vector, which is the penultimate layer in a DL model, is transformed using a transition matrix T by formula (6) to produce results that are understandable to the end user.

Suppose there is an expert in the subject area of the problem under consideration (i.e., the end-user) who compiles an exhaustive list of features by which they determine the belonging of an object to a particular class. Further, for each feature from the list of features, the expert indicates the numerical intervals into which the value of the feature should fall for the classes under consideration. Finally, for each instance (object) from the training dataset, the value of each feature is calculated.

The values of features can be determined in several ways, namely:

- Empirically, using the expert's knowledge of the subject area of the problem under consideration.
- Using formulas or statistical indicators that are understandable to the end user.
- By visual representation (in various ways) of a fragment of a signal or image, in comparison with similar fragments from labeled training data.
- Utilizing visual analytics.
- Using ML models specially built for this case.
- Using DL models specially built for this case.
- In other ways.

Figure 3 shows the main steps of the method for constructing matrix B, according to the proposed scalable approach.

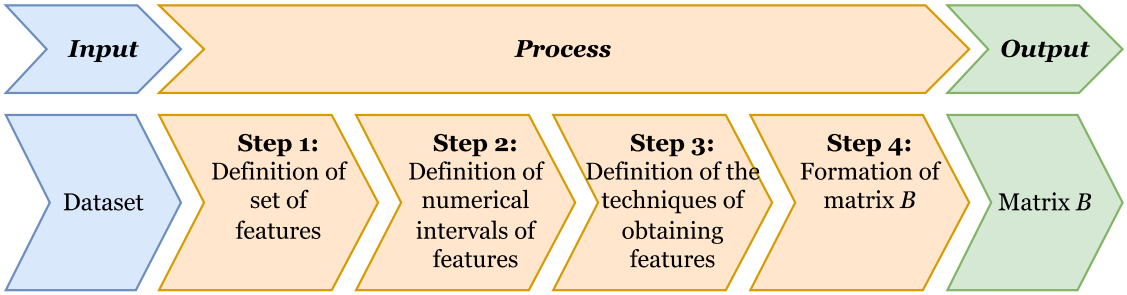


Figure 3: Diagram of constructing matrix B for the proposed scalable approach of visual analytics with a transition matrix from the DL model to the so-called feature model.

Below we provide the following steps to build matrix B.

Input information: An expert in the subject area of the problem under consideration and the dataset on which the DL model was trained.

Step 1: The expert compiles a list of features, by the values of which the end user sees the belonging of the object to a certain class. The list should be as exhaustive as possible and, if possible, exclude mutual intersections between classes, that is, the belonging of an object to several classes simultaneously.

Step 2: For each feature defined in the previous step, numerical intervals of belonging and non-belonging to the class are indicated.

Step 3: The expert defines ways to obtain numerical values of the feature: empirically, by calculation formulas, by visual representation of the current fragment in comparison with similar fragments from labeled training data, using visual analytics tools, by created ML or DL models for this case, or by other methods.

Step 4: For each object from the training set, features are determined, and the corresponding row of matrix B is formed.

Output information: Matrix *B*.–

3.3. Evaluation criteria

In this work, Cohen's Kappa coefficient (κ) is used to evaluate the quality of the proposed approach. The κ coefficient is a reliable statistical indicator for evaluating inter-expert reliability for qualitative (categorical) elements. It quantifies the level of agreement between two experts beyond chance.

The formula for Cohen's Kappa coefficient κ is as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{9}$$

where P_o is the level of observed (empirical) agreement between two experts, and P_e is the level of expected (calculated) agreement between the same experts.

For the problem of binary classification of medical signals and/or images with a confusion matrix consisting of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), the elements of formula (9) have the following form:

$$P_o = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}. \tag{10}$$

$$P_e = \left(\frac{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) + (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}{(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})^2}\right). \tag{11}$$

In the formula (10), P_o is the proportion of cases in which the DL model and the human expert come to a consistent decision. Instead, in the formula (11), P_e is calculated based on the marginal sum values of the decisions of the two experts.

The value of the κ coefficient according to formula (9) is in the range [−1;1], where 1 denotes perfect agreement, 0 stands for no agreement, and negative values reflect less agreement than expected by chance.

4. Results and discussion

Experimental evaluation of the proposed scalable approach was performed by solving two problems with DL models:

- Detection of pathologies of heart activity disorders (arrhythmias) based on electrocardiogram (ECG) signals [26].
- Detection of pathologies in the areas of the myocardium, left and right ventricles, based on magnetic resonance imaging (MRI) [27].

Next, we will present the results and discussion of the application of the proposed approach to explaining the decisions made by DL models.

4.1. Detection of pathologies of heart activity based on ECG signals

The proposed approach was validated by the constructed DL_{ECG} model for the problem of detecting pathologies of heart activity (arrhythmias) based on ECG signals [26]. Below we will describe the training dataset, the DL_{ECG} model, and the set of features that explained the decisions and results of the proposed approach (the value of κ).

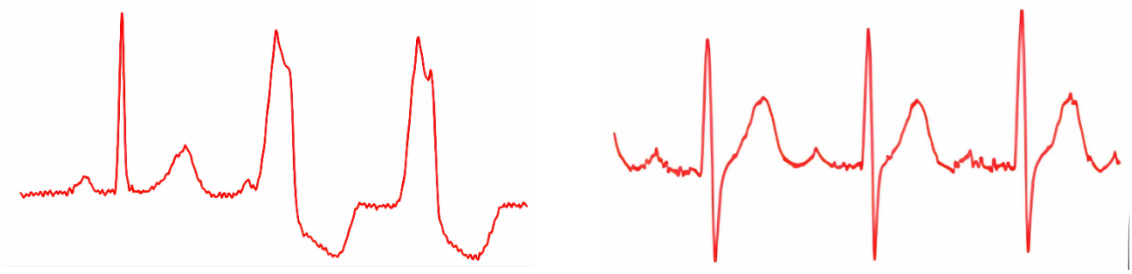
4.1.1. Training dataset and DL model

The problem of detecting pathologies of heart activity (arrhythmias) based on ECG signals was solved using the reference dataset MIT-BIH Arrhythmia Database (MIT-BIH) [28]. The training of the DL_{ECG} model was performed on 80% of the data from MIT-BIH. Given the annotations of the MIT-BIH set, the following classes/pathologies were selected for the classification problem:

1. Normal beat.

2. Premature ventricular contraction.
3. Paced beat.
4. Right bundle branch block beat.
5. Left bundle branch block beat.
6. Atrial premature beat.
7. Fusion of ventricular and normal beat.
8. Fusion of paced and normal beat.
9. Others.

Input information for training and testing the DL_{ECG} model is presented as a triad of cardiac cycles – in the center is the main cardiac cycle, to which the previous and next cardiac cycles were added (Figure 4).



(a) **(b)**

Figure 4: Examples of presenting the input fragment of the ECG signal as a triad of cardiac cycles: (a) with a heterogeneous display of the R-peak and (b) with a homogeneous expression of the R-peak.

In this work, the DL_{ECG} model was created based on the modified architecture from our previous work [26]. The classification accuracy for the training set was 99.95%, for the test set – 99.13%.

The penultimate layer of the DL_{ECG} model contained 8192 neurons, and the number of samples in the training sample was 52,180. Accordingly, the size of the A_{ECG} matrix was m_{ECG} = 52,180 – the number of objects from the training subsample of the MIT-BH dataset, k_{ECG} = 8,192 – the number of features formed by the DL_{ECG} model.

4.1.2. Features for explanation

For the experiment, the class identified as «Premature Ventricular Contraction» (PVC) or «Ventricular Extrasystole» was selected. Cardiologists pinpointed key ECG features to identify this condition:

- Absence of the P-peak.
- An expanded and deformed QRS complex, with deformation indicating a change in shape. Specifically, a right ventricular extrasystole mimics a left bundle branch block in lead V1, while a left ventricular extrasystole resembles a right bundle branch block.
- A complete compensatory pause, defined as the interval between two sinus rhythm ventricular complexes surrounding an extrasystole, which equals double the RR interval of the sinus rhythm. This pause marks the time from the extrasystole until the next normal contraction.

4.1.3. Statistical analysis for ECG classification

Given the significant amount of the training set and the significant time of experts regarding filling in the values of features, non-empirical methods of determining the value of features were used in this work.

For the «Absent P-peak» feature, the visual analytics tool Principal Component Analysis (PCA) was used. The application of PCA and the reduction of data dimension to 3 made it possible to make

sure that the signal fragment with the presence and absence of P-peaks is separate. Given this, the presence/absence of the P-peak was determined using the Neurokit2 toolkit [29]. Visualization of dimensionality reduction by PCA for the «Absent P-peak» feature is shown in Figure 5.

For the «Expanded and deformed QRS complex» feature, due to the complexity of its detection by other methods, it is proposed to use a specially trained neural network.

The «Full compensatory pause» feature. A compensatory pause is the time elapsed after an extrasystole until the occurrence of a normal contraction. Therefore, in the case when the extrasystole is located between other extrasystoles, this calculation is not performed and is calculated only for the last case of extrasystole in the sequence.

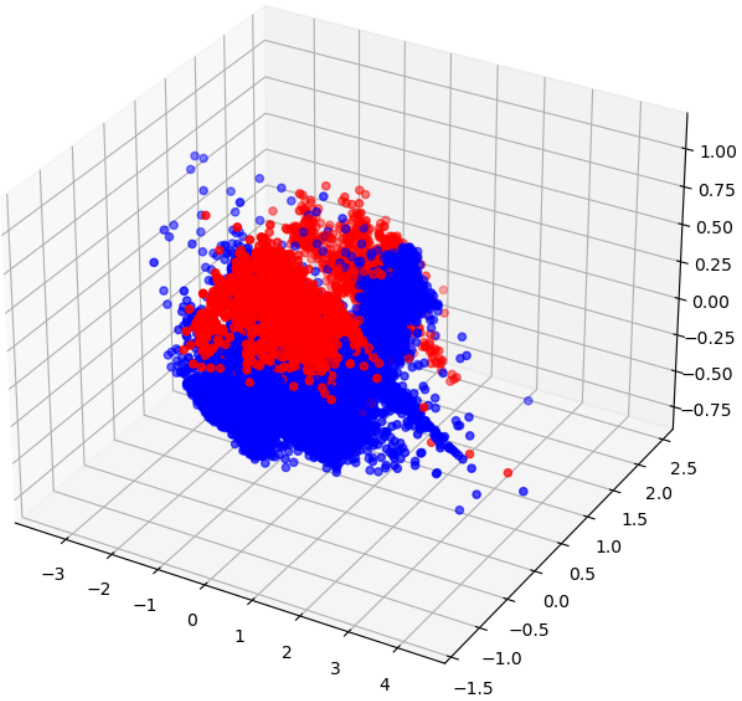


Figure 5: The result of applying PCA to determine the «Absent P-peak» feature.

The presence or absence of this feature was checked as follows:

1. Using the Neurokit2 package, the average R-R interval between normal cardiac cycles (RR_n) was determined.
2. The interval between the R-peak with extrasystole and the R-peak of the previous cycle (RR_{prev}) was determined.
3. The interval between the R-peak with extrasystole and the R-peak of the next normal cycle (RR_{next}t) was determined.
4. A full compensatory pause was determined under the following condition:

$$RR_{prev} + RR_{next} = 2 \times RR_n,$$

$$\left(2 \times RR_n - \left(RR_{prev} + RR_{next}\right)\right) < \text{tolerance}.$$

According to the above rules, the values of features were determined for each sample from the training set, and, in this way, matrix B was obtained. Further, according to formula (6), the transition matrix T was determined.

- Coefficient κ_1 was calculated to evaluate the agreement between the class annotations in the test set and the class predictions made by the DL_{ECG} model.
- Coefficient $\kappa_2$was calculated to determine the agreement between the class annotations obtained by the DL_{ECG} model and those obtained by the approximated feature values.

The resulting κ_1 was 0.98. To assess the precision of this estimate, a 95% confidence interval (CI) was computed, resulting in a CI of 0.96 to 1.00. Additionally, the associated p-value was calculated to be <0.001, indicating that the observed agreement is highly unlikely to be due to chance. The high

κ_1 value, combined with the narrow confidence interval, signifies an almost perfect agreement between the expert annotations and the DL_{ECG} model's predictions. This strong agreement is further supported by the p-value, which confirms the statistical significance of the result.

The resulting κ_2 was 0.89, with a 95% CI of 0.85 to 0.93, and a p-value of <0.001. This κ_2 value indicates a strong agreement between the model's predictions and the approximated feature-based annotations, although it is slightly lower than κ_1. The slightly broader confidence interval reflects a bit more variability in the agreement, which could be attributed to the approximation process. Nonetheless, the p-value still indicates that this agreement is highly significant, and the κ_2 value demonstrates that the expert's features can reliably replicate the decisions made by the DL_{ECG} model.

The comparison between κ_1 and κ_2, along with their respective confidence intervals and p-values, provide a comprehensive understanding of the model's performance. The high κ_1 value, coupled with a narrow confidence interval and a significant p-value, confirms the DL_{ECG} model's capability to accurately classify ECG signals in alignment with expert annotations. The slightly lower, yet still strong, κ_2 value suggests that while the approximated feature values can effectively mirror the model's decisions, there is a slight decrease in agreement, which may warrant further investigation into the approximation methods, or the features used.

4.2. Detection of pathologies of heart activity based on MRI

The proposed scalable approach was also validated by the DL_{MRI} model for the problem of detecting pathologies of heart activity based on MRIs [27].

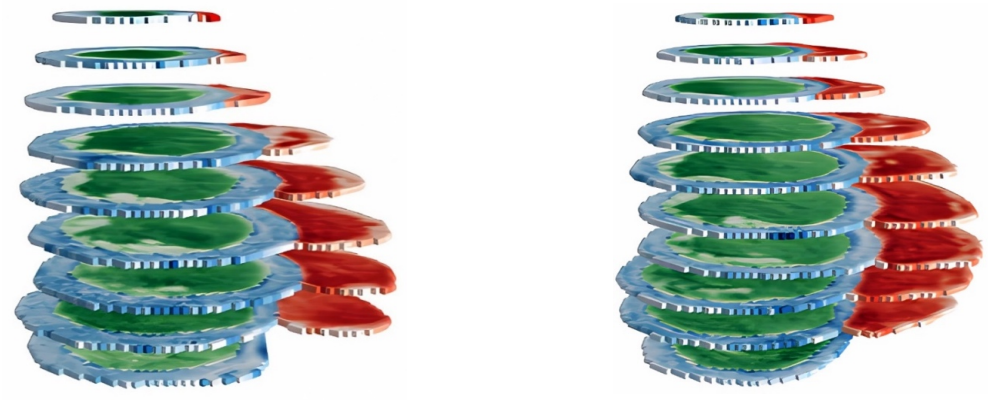
Next, we will briefly describe the training dataset of MRIs, the DL_{MRI} model, the set of features that explained the decisions, and the results of the proposed approach (the value of κ).

4.2.1. Training dataset and DL model

For the problem of detecting pathologies of heart activity based on MRIs, a modified dataset of the Automatic Cardiac Diagnosis Challenge (ACDC) [30] was used. Samples of the ACDC set of 100 and 50 patients were used for training and testing the network, respectively. Given the annotations to the ACDC set, the following classes/pathologies were selected for classification:

- Normal condition.
- Dilated cardiomyopathy (DCM).
- Hypertrophic cardiomyopathy (HCM).
- Myocarditis (MINF).
- Arrhythmogenic right ventricular cardiomyopathy (ARV).

An example of presenting input data to the DL model according to the ACDC dataset is illustrated in Figure 6.



(a) **(b)**

Figure 6: Visualization of input data for classification: segmented MRIs for ES phase (a) and ED phase (b) with preservation of MRI signal intensity values.

The DL_{MRI} model was created based on the modified architecture from our work [27]. The classification accuracy for the test set of MRIs was over 96.5%. The size of the A_{MRI} matrix was m_{MRI} = 100 – the number of objects from the training subsample of the ACDC dataset, k_{MRI} = 1024 – the number of features formed by the DL_{MRI} model.

4.2.2. Features for explanation

In modern medical practice, various features from MRIs are actively used for the classification of heart diseases, which correspond to the classes considered above. For our task, 20 features were used, selected by experts (cardiologists). At the same time, according to the features of identifying pathologies identified by the doctor, the following set of geometric features was formed for further classification.

1. The ratio of the volume of the left ventricle to the volume of the right ventricle at the end of systole.
2. The volume of the left ventricle at the end of systole.
3. The ratio of the volume of the left ventricle to the volume of the right ventricle at end-diastole.
4. The volume of the left ventricle at end-diastole.
5. The volume of the right ventricle at end-systole.
6. The volume of the right ventricle at end-diastole.
7. Ejection fraction of the left ventricle.
8. Ejection fraction of the right ventricle.
9. The ratio of myocardial volume to left ventricular volume at the end of systole.
10. Myocardial mass at the end of diastole.
11. Myocardial volume at the end of systole
12. The ratio of myocardial mass to left ventricular volume at the end of diastole.
13. Maximum average myocardial wall thickness at end-diastole.
14. Maximum average myocardial wall thickness at end systole.
15. Mean standard deviation of myocardial wall thickness at end-systole.
16. Mean standard deviation of myocardial wall thickness at end-diastole.
17. Standard deviation of the standard deviation of myocardial wall thickness at end-diastole.
18. Standard deviation of the standard deviation of myocardial wall thickness at end-systole.
19. Standard deviation of mean myocardial wall thickness at end-diastole.
20. Standard deviation of mean myocardial wall thickness at end-systole.

4.2.3. Statistical analysis for MRI classification

For further experiments, the class «DCM was chosen. Since the volume of the training set was small, it did not take much time for experts to fill in the values of features. Cardiologists as experts determined the values of features for each sample from the training set and, in this way, matrix B was formed. Further, according to formula (6), the transition matrix T was determined.

For each object from the test set, the values of features were approximated according to the formula (8). Subsequently, two key values of κ were also calculated to assess the agreement between the predictions DL_{MRI} of and the expert annotations. These κ values provide a quantitative measure of the reliability and consistency of the DL_{MRI} model's performance, and they were computed using the formula (9).

- Coefficient κ_1 was calculated to evaluate the agreement between the class annotations in the test set and the classifications made by the DL_{MRI} model.
- Coefficient κ_2 was calculated to determine the agreement between the class annotations obtained by the DL_{MRI} model and those obtained through the approximated feature values.

The resulting κ_1 value was 0.87, indicating a very good level of agreement. To further substantiate this finding, a 95% confidence interval was computed, yielding a range of 0.83 to 0.91. Additionally, the associated p-value was determined to be <0.001, confirming the statistical significance of this agreement. The relatively narrow confidence interval suggests that the κ_1 value is a stable estimate, and the low p-value strongly supports that the observed agreement is unlikely due to random chance. This high κ_1 value aligns well with the model's overall performance, highlighting the DL_{MRI} model's robust capability in accurately classifying MRI data per expert labels.

The resulting κ_2 value was 0.80, with a 95% confidence interval of 0.76 to 0.84, and a p-value of <0.001. This κ_2 value indicates a significant match, though slightly lower than κ_1, reflecting a strong agreement but with slightly more variability. The broader confidence interval compared to κ_1 suggests some degree of uncertainty, possibly arising from the approximation process or the inherent variability in the feature values. Nonetheless, the significant p-value still confirms that this agreement is meaningful and not due to random variation.

The comparison between κ_1 and κ_2, along with their respective confidence intervals and p-values, provides a detailed insight into the DL_{MRI} model's performance and the reliability of the feature approximation approach. The κ_2 value of 0.87, with a narrow confidence interval and a highly significant p-value, underscores the strong alignment between the model's predictions and expert annotations. Meanwhile, the slightly lower κ_2 value of 0.80 suggests that while the approximation method is effective, there is a minor decrease in agreement that could be attributed to the complexity of the MRI data or the approximation method itself.

Overall, the inclusion of these detailed statistical indicators, i.e., κ values, confidence intervals, and p-values, adds robustness to the analysis, strengthening the reliability and validity of both DL_{MRI} and DL_{ECG} model's performance and the transparency of the proposed scalable approach.

4.3. Limitations of the proposed scalable approach

The proposed scalable approach of visual analytics has shown high reliability, with strong agreement between the DL_{ECG} and DL_{MRI} models and expert annotations, demonstrating excellent interpretability. However, several limitations must be noted:

- Feature selection bias. Relying heavily on expert knowledge for feature selection can lead to biases, which may result in overfitting and reduce the model's ability to generalize to unseen data. This risk is higher when features are tailored to specific datasets without sufficient variability.
- Complexity versus interpretability. As DL models become more complex, maintaining interpretability becomes increasingly challenging, especially when handling large, interconnected features. There is a trade-off between performance and explainability that must be carefully managed to prevent overwhelming users with overly technical or unintuitive explanations.
- Inconsistent feature quantification. Experts may struggle to consistently assign numerical values to features, which can lead to variability in training data and affect the reliability of the model's explanations, particularly in cases where precise quantification is needed.
- Adaptability to different datasets. The scalability and adaptability of this approach to different datasets and clinical settings remain uncertain. Significant reconfiguration may be required for new data sources, limiting the approach's broad applicability.

To overcome these limitations, future research should focus on reducing biases in feature selection, ensuring models can generalize to unseen data, refining feature quantification methods, and improving the scalability and adaptability of the approach across different medical datasets and environments.

4.4. Discussion on future work

First, while the proposed approach is scalable, its adaptability to diverse medical datasets and clinical environments requires further exploration. Ensuring that the approach can generalize effectively across various types of medical data, such as radiological images, electrocardiograms, and clinical reports, is crucial for its widespread applicability. Each dataset presents unique challenges, ranging from differences in imaging modalities to variations in clinical practices, that may affect the model's performance and interpretability. Testing this approach across multiple datasets, including those with varying quality, data distributions, and medical conditions, will help establish its robustness and ensure that it can be deployed effectively in real-world healthcare settings. In addition, clinical environments differ significantly, and models may require fine-tuning to adapt to specific clinical workflows or diagnostic guidelines. Therefore, future work should focus on evaluating the model's performance in diverse clinical scenarios, ensuring that its scalability is matched by a capacity to generalize effectively across both datasets and environments

Integrating domain-specific knowledge, such as clinical guidelines, expert rules, or established diagnostic protocols, is essential to improving both the quality of explanations and the performance of the model. By embedding clinical knowledge directly into the feature selection process, the model can generate more contextual explanations that resonate with healthcare professionals. For instance, using expert-defined diagnostic rules for conditions like arrhythmia or cardiomyopathies ensures that the features highlighted by the model are not only relevant but also clinically actionable. This approach might enhance interpretability and strengthen the trustworthiness of the AI system, as clinicians can better align the model's predictions with their own clinical judgments. In this regard, future research might also focus on building frameworks that allow the seamless integration of such domain knowledge into AI systems, ensuring that they remain transparent while adhering to medical best practices.

Finally, beyond technical performance, ethical considerations must be at the forefront when deploying explainable AI in healthcare. The introduction of AI-generated explanations in clinical settings can have profound impacts on decision-making processes and patient outcomes. Transparency is crucial, but it is equally important to ensure that these explanations do not lead to a false sense of security or over-reliance on AI systems. Healthcare providers must remain accountable for final decisions, particularly when AI-generated insights conflict with clinical intuition. Moreover, the use of AI in sensitive scenarios, such as life-threatening conditions, raises questions about the adequacy of AI-driven interventions and their potential to introduce bias or errors into the healthcare process. The integration of explainable AI thus necessitates the development of ethical frameworks that guide its application, ensuring that patient autonomy, informed consent, and equitable care are robustly preserved.

5. Conclusions

In this study, we introduced a scalable approach designed to make DL model decisions more explainable by mapping them to features that are understandable to healthcare experts. The approach was rigorously tested on two distinct medical datasets: ECG signals used for detecting arrhythmias and MRI scans for classifying heart diseases. In the ECG experiment, the DL model achieved a Cohen's Kappa coefficient of 0.89, demonstrating a nearly perfect agreement with expert annotations. Similarly, for the MRI dataset, the model showed a Cohen's Kappa coefficient of 0.80, indicating a strong agreement, thereby underscoring the reliability of the proposed method in replicating expert-level decisions. Overall, the obtained results underscore the approach's effectiveness in enhancing model explainability. Nonetheless, the approach faces challenges, particularly in creating an exhaustive feature list and ensuring consistent numerical feature values.

Future work should focus on improving the feature selection process by incorporating more standardized and automated methods to reduce variability.

References

[1] R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, A historical perspective of explainable artificial intelligence, WIREs Data Min. Knowl. Discov. 11.1 (2020) e1391. doi:10.1002/widm.1391.

[2] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, A. Hussain, Interpreting black-box models: A review on explainable artificial intelligence, Cogn. Comput. 16.1 (2024) 45–74. doi:10.1007/s12559-023-10179-8.

[3] K. Vredenburgh, The right to explanation, J. Political Philos. 30.2 (2022) 209–229. doi:10.1111/jopp.12262.

[4] V. K. Venkatesan, M. T. Ramakrishna, I. Izonin, R. Tkachenko, M. Havryliuk, Efficient data preprocessing with ensemble machine learning technique for the early detection of chronic kidney disease, Appl. Sci. 13.5 (2023) 2885. doi:10.3390/app13052885.

[5] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, et al., Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions, Inf. Fusion 106 (2024) 102301. doi:10.1016/j.inffus.2024.102301.

[6] United States: Commerce Department: National Institute of Standards and Technology (NIST), P. J. Phillips, P. J. Phillips, P. J. Phillips, C. A. Hahn, P. C. Fontana, A. N. Yates, K. Greene, D. A. Broniatowski, M. A. Przybocki, et al., Four principles of explainable artificial intelligence, 8312, Commerce Department, Gaithersburg, MD, USA, 2021. doi:10.6028/NIST.IR.8312.

[7] J. Pääkkönen, P. Ylikoski, Humanistic interpretation and machine learning, Synthese 199.1 (2021) 1461–1497. doi:10.1007/s11229-020-02806-w.

[8] T. Räuker, A. Ho, S. Casper, D. Hadfield-Menell, Toward transparent AI: A survey on interpreting the inner structures of deep neural networks, in: 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), IEEE, Inc., New York, NY, USA, 2023, pp. 464–483. doi:10.1109/SaTML54575.2023.00039.

[9] A. Notovich, H. Chalutz-Ben Gal, I. Ben-Gal, Explainable artificial intelligence (XAI): Motivation, terminology, and taxonomy, in: L. Rokach, O. Maimon, E. Shmueli (Eds.), Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook, Springer International Publishing, Cham, 2023, pp. 971–985. doi:10.1007/978-3-031-24628-9_41.

[10] D. Kim, Y. Song, S. Kim, S. Lee, Y. Wu, J. Shin, D. Lee, How should the results of artificial intelligence be explained to users? - Research on consumer preferences in user-centered explainable artificial intelligence, Technol. Forecast. Soc. Chang. 188 (2023) 122343. doi:10.1016/j.techfore.2023.122343.

[11] Y. Wang, S. H. Chung, Artificial intelligence in safety-critical systems: A systematic review, Ind. Manag. & Data Syst. 122.2 (2021) 442–470. doi:10.1108/IMDS-07-2021-0419.

[12] M. Mora-Cantallops, E. García-Barriocanal, M.-Á. Sicilia, Trustworthy AI guidelines in biomedical decision-making applications: A scoping review, Big Data Cogn. Comput. 8.7 (2024) 73. doi:10.3390/bdcc8070073.

[13] P. Radiuk, O. Barmak, I. Krak, An approach to early diagnosis of pneumonia on individual radiographs based on the CNN information technology, Open Bioinform. J. 14.1 (2021) 92–105. doi:10.2174/1875036202114010093.

[14] M. L. Smith, L. N. Smith, M. F. Hansen, The quiet revolution in machine vision - A state-of-the-art survey paper, including historical review, perspectives, and future directions, Comput. Ind. 130 (2021) 103472. doi:10.1016/j.compind.2021.103472.

[15] D. Khurana, A. Koli, K. Khatter, S. Singh, Natural language processing: State of the art, current trends and challenges, Multimedia Tools Appl. 82.3 (2023) 3713–3744. doi:10.1007/s11042-022-13428-4.

[16] P. Radiuk, O. Barmak, E. Manziuk, I. Krak, Explainable deep learning: A visual analytics approach with transition matrices, Mathematics 12.7 (2024) 1024. doi:10.3390/math12071024.

[17] R. Bassiouny, A. Mohamed, K. Umapathy, N. Khan, An interpretable object detection-based model for the diagnosis of neonatal lung diseases using ultrasound images, in: 2021 43rd Annual

International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, Inc., New York, NY, USA, 2021, pp. 3029–3034. doi:10.1109/EMBC46164.2021.9630169.

[18] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, P. Lambin, Transparency of deep neural networks for medical image analysis: A review of interpretability methods, Comput. Biol. Med. 140 (2022) 105111. doi:10.1016/j.compbiomed.2021.105111.

[19] M.-C. Chan, K.-C. Pai, S.-A. Su, M.-S. Wang, C.-L. Wu, W.-C. Chao, Explainable machine learning to predict long-term mortality in critically ill ventilated patients: A retrospective study in central Taiwan, BMC Med. Inform. Decis. Mak. 22.1 (2022) 75. doi:10.1186/s12911-022-01817-6.

[20] 20. K. Lu, Y. Tong, S. Yu, Y. Lin, Y. Yang, H. Xu, Y. Li, S. Yu, Building a trustworthy AI differential diagnosis application for Crohn's disease and intestinal tuberculosis, BMC Med. Inform. Decis. Mak. 23.1 (2023) 160. doi:10.1186/s12911-023-02257-6.

[21] P. A. Moreno-Sánchez, Improvement of a prediction model for heart failure survival through explainable artificial intelligence, Front. Cardiovasc. Med. 10 (2023) 1219586. doi:10.3389/fcvm.2023.1219586.

[22] E. Pintelas, I. E. Livieris, P. Pintelas, explainable feature extraction and prediction framework for 3D image recognition applied to pneumonia detection, Electronics 12.12 (2023) 2663. doi:10.3390/electronics12122663.

[23] D. S. Cvetković Ilić, Y. Wei, Completions of operator matrices and generalized inverses, in: D. S. Cvetković-Ilić, Y. Wei (Eds.), Algebraic Properties of Generalized Inverses, Springer, Singapore, 2017, pp. 51–88. doi:10.1007/978-981-10-6349-7_3.

[24] D. Kalman, A singularly valuable decomposition: The SVD of a matrix, Coll. Math. J. 27.1 (2002) 2–23. doi:10.2307/2687269.

[25] Krak, I., Barmak, O., Manziuk, E., Kulias, A. Data classification based on the features reduction and piecewise linear separation. (2020) Advances in Intelligent Systems and Computing, 1072, pp. 282-289. doi: 10.1007/978-3-030-33585-4_28.

[26] O. Kovalchuk, O. Barmak, P. Radiuk, I. Krak, ECG arrhythmia classification and interpretation using convolutional networks for intelligent IoT healthcare system, in: 1st International Workshop on Intelligent & CyberPhysical Systems (ICyberPhyS-2024), CEUR-WS.org, Aachen, 2024, pp. 47–62. URL: https://ceur-ws.org/Vol-3736/paper4.pdf.

[27] V. Slobodzian, P. Radiuk, A. Zingailo, O. Barmak, I. Krak, Myocardium segmentation using two-step deep learning with smoothed masks by Gaussian blur, in: 6th International Conference on Informatics & Data-Driven Medicine (IDDM-2023), CEUR-WS.org, Aachen, 2024, pp. 77–91. URL: https://ceur-ws.org/Vol-3609/paper7.pdf.

[28] G. B. Moody, R. G. Mark, MIT-BIH arrhythmia database, Software, v. 1.0.0, Data Collection, physionet.org, 2005. doi:10.13026/C2F305.

[29] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, S. H. A. Chen, NeuroKit2: A Python toolbox for neurophysiological signal processing, Behav. Res. Methods 53.4 (2021) 1689–1696. doi:10.3758/s13428-020-01516-y.

[30] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?, IEEE Trans. Med. Imaging 37.11 (2018) 2514–2525. doi:10.1109/TMI.2018.2837502.