Exploring CNN and XAI-based Approaches for Accountable MI Detection in the Context of IoT-enabled Emergency Communication Systems

Helene Knof*
helene.knof@fokus.fraunhofer.de
Fraunhofer Institute for Open Communication Systems (FOKUS)

Prachi Bagave*
p.bagave@tudelft.nl
Delft University of Technology, The Netherlands

Michell Boerger
michell.boerger@fokus.fraunhofer.de
Fraunhofer Institute for Open Communication Systems (FOKUS)

Nikolay Tcholtchev
nikolay.tcholtchev@fokus.fraunhofer.de
Fraunhofer Institute for Open Communication Systems (FOKUS)

Aaron Yi Ding
Aaron.Ding@tudelft.nl
Delft University of Technology, The Netherlands

Abstract

The ageing European population and the expected increasing number of medical emergencies put pressure on the medical sector and existing emergency infrastructures, which calls for new innovative digital solutions. In parallel, the increasing utilization of the Internet of Things (IoT) has enabled the collection of real-time data, allowing for the autonomous detection of acute medical emergencies. In this context, this paper presents two distinct machine learning (ML) models that leverage sensor data to autonomously detect emergencies. These models are intended to be integrated into an IoT-enabled next-generation emergency communications system (NG112) capable of detecting emergencies, initiating emergency calls (eCalls), and providing relevant information to emergency call takers, which reduces response time. Thereby, this paper focuses on the accountable detection of myocardial infarctions (MIs), commonly known as heart attacks, based on electrocardiogram (ECG) data. To realize this, two disparate models working on fundamentally different data structures are proposed and compared: A one-dimensional convolutional neural network (CNN) operating on the raw ECG signals and a GoogLeNet-based model trained on ECG images. The PTB-XL dataset is used to evaluate the proposed models, and the results indicate the 1D CNN exhibits a favourable trade-off between precision and recall for the eCall use case. Finally, the paper also discusses applying eXplainable AI (XAI) methods to achieve explainability for the ML models, paving the way for an accountable and reliable implementation in safety-critical systems.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Applied computing** → *Health informatics*; • **Social and professional topics** → Medical technologies.

Keywords: machine learning, emergency detection, myocardial infarctions, explainable AI, datasets, neural networks

*Both authors contributed equally to this research.

1 Introduction

The pervasive and consistently increasing adoption of the Internet of Things has enabled the collection of large amounts of real-time data at the Edge. Environmental and near-body sensors such as smartwatches, accelerometers, or wearable ECG sensors facilitate accumulating and monitoring an accurate and temporally updated representation of users' ambient conditions and body vitals. This creates the opportunity to leverage the available IoT data to automatically detect patterns and anomalies that may indicate emergencies such as fires, falls, or even acute cardiac issues.

At the same time, the constantly ageing European society [6] poses challenges to the medical system and emergency infrastructure, and calls for innovative and automated digital solutions. Precisely, the prompt detection of medical emergencies and immediate initiation of medical assistance are becoming increasingly important in light of the rising number of elderly patients [6]. In this context, the advancing digitalization and the increased availability of IoT sensor data can significantly contribute to the automated detection of emergencies leading to an early response.

Therefore, we present and compare two disparate approaches for realising ML-based models that leverage sensor data to automatically detect acute medical emergencies in this paper. We investigate these models in the context of the use case of an IoT-enabled and IP-based next-generation emergency communications system. Precisely, we aim to integrate the presented models into an existing NG112 system that is capable of (1) automatically detecting emergencies at the patient side based on IoT sensor data, (2) automatically initiating emergency calls in case emergencies are detected, and (3) providing the emergency call taker with relevant information and reasoning behind the decision making.

In this paper, we will address the automated detection of myocardial infarctions exemplifying a time-critical emergency. Indeed, cardiovascular diseases, including MIs, are the leading cause of death worldwide [20], and early detection

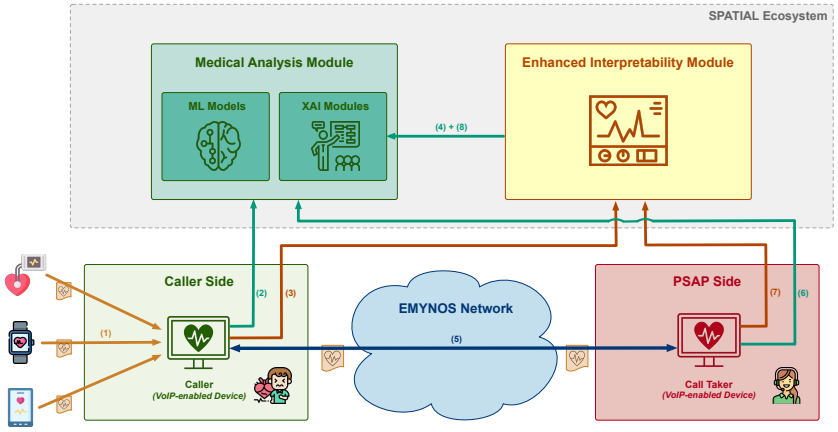

Figure 1. Integration of the ML-based emergency detection into the NG112 emergency e-calling system EMYNOS.

is crucial in reducing mortality rates. MIs can be diagnosed with ECGs, which measure the heart's electrical activity.

Given that the automated ML-based detection of MIs and the autonomous establishment of emergency calls represent a safety-critical application, the accountability of the discussed ML-based system is of utmost importance. Explanations of AI systems can help enhance the accountability of the system if they are interpretable to the decision makers [3]. We will therefore give an outlook on how to apply and compare different explainable AI (XAI) methods to the proposed MI detection models. This is also the motivation for the two disparate methods discussed in our work. Our first method uses the ECG as a waveform time series data. However, due to the availability of a number of XAI methods for image classifiers, in our second method, we use ECGs as images for MI detection. The envisioned results are supposed to provide the patient and emergency call taker with reasonable explanations about decisions of the ML models.

In summary, we contribute to the field of accountable ML in safety-critical systems by:

- Proposing and comparing two ML models for MI detection on ECG data in IoT-based NG112 systems,
- Demonstrating the application of XAI methods to enhance the models' accountability,
- Recommending future research directions for evaluating the accountability of the proposed models.

To achieve this, we start with presenting the NG112 system and related work in Section 2, followed by a description of the methodology (Section 3) and evaluation results (Section 4). An outlook on applying XAI methods is provided in Section 5, followed by a conclusion in Section 5.

2 Background

In this section, we discuss the NG112 emergency communication system which serves as the application basis for the work provided in this paper, and also highlight the most relevant work in the context of MI detection using ML.

2.1 The NG112 emergency communication system

For the above-mentioned IoT-enabled emergency communication system, we build on the results of the EU-funded H2020 project EMYNOS (nExt generation eMergencY commuNicatiOnS)[1]. In the course of this project, Rebahi et al. [18] specified an NG112 emergency communication system that allows (1) to integrate various eHealth sensors to implement appropriate monitoring and (2) to establish VoIP-based emergency calls that provide the functionality to transmit sensor data between the caller and callee. Figure 1 visualizes a simplified version of a corresponding testbed of the EMYNOS framework, which was developed and tested by Barakat et al. [4] and Kumar Subudhi et al. [12].

We aim to enhance the EMYNOS platform by implementing an AI system that accurately recognizes emergencies and autonomously initiates eCalls while being accountable and effective. In addition to the automated emergency recognition, we want to provide XAI explanations to the patient and emergency call taker. These should help the latter better assess the emergency situation, estimate necessary medical resources, instruct potentially available first responders, and even recognize and reject hoax calls.

To realize the automated eCall functionality, we will integrate the ML models and XAI functionality proposed in this paper into the EMYNOS platform. As depicted in Figure 1, the *Medical Analysis Module* (MAM) provides access to the models and XAI methods. When a user at the Caller Side receives new ECG readings from connected IoT devices, the data is forwarded to the MAM. The MAM then analyzes the ECG, provides a classification score, and delivers XAI explanations to the patient. If an emergency is detected, an

[1]H2020 EMYNOS: https://www.emynos.eu/, as of date 27.07.2023

eCall is automatically initiated, and the sensor data is transmitted to the emergency call taker, who can also utilize the MAM to obtain relevant explanations. Additionally, we will grant access to the *Enhanced Interpretablity Module*, a WebApp that offers an interactive GUI interface to enhance the interpretability of the explanations provided.

2.2 Related Work

Recently, there have been a growing number of studies for the detection of heart-related diseases from ECG data. The two most commonly used open ECG databases are called PTB [8] and PTB-XL [28]. The two databases are disjoint, with PTB-XL being larger and with broader annotations. Some of these works also extend their work from the MI detection towards explaining their models for increased transparency.

As the PTB-XL dataset is relatively new, there are only a few studies using it in their work. Hammad et al. [9] uses a CNN combined with an SVM for the classification of four cardiovascular pathologies available in the PTB-XL dataset, and achieved an accuracy of 98.9%. A similar study by Chen et al. [5] used the PTB-XL dataset for training and validation, but used a dataset from Chapman university and Shaoxing People's Hospital [30] for testing. They used a residual network for MI detection and achieved an AUC of 97.7%, specificity of 95.1%, and sensitivity of 95.1%. Another study by Martin et al. [16] preprocessed the signal by using single lead information aligned in time, and achieved an accuracy of around 84.1%, by using deep LSTM. On the other hand, Ma et al. [14] focused on simplifying the model to reduce the prediction time of MI, and used a convolutional dendrite net, achieving a decent accuracy of 96.80%. Contrary to all these studies with time series data as the input data type, Fang et al. [7] proposed a novel method by generating 3D images with the 12-lead ECG data, and and achieved an accuracy of 97.23% by training a multi-VGG deep neural network. They also explained their AI outputs by using Grad-CAM++ on these images. Anand et al. [1], on the other hand used variations of Spatio-Temporal CNNs, and used SHAP to explain their model along with the medically relevant information.

In addition, the studies using the PTB database achieved even better performance results. Zhang et al. [29] extracted single heartbeats for MI detection and MI localization, and achieved an accuracy, sensitivity, and specificity of 99.88%, 99.98% and 99.39% respectively. Rai and Chatterjee [17] and Han and Shi [10] compared deep learning methods like CNN, hybrid CNN-LSTM, and ensemble techniques for MI detection and achieved accuracy as high as 99.8% [17]. In addition to the CNN architecture, Jahmunah et al. [11] also used LRP for explanations and Strodthoff and Strodthoff [24] used Grad-CAM.

A few other studies used ECG-plotted images from the PTB dataset as input data and used CNN for classification Makimoto et al. [15], Uchiyama et al. [27]. [15] also used Grad-CAM for explanations.

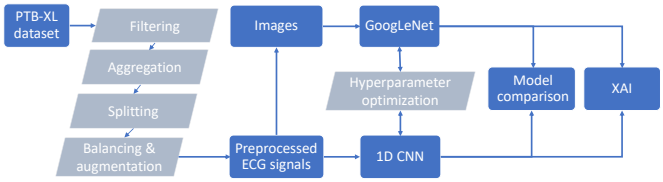

Figure 2. Overview of the methodology pipeline.

3 Methodology for ML-based MI Detection

Figure 2 shows an overview of our approach to obtain and compare two disparate ML models capable of detecting indications of MIs in provided ECG input data. As motivated in Section 1, we compare two models that operate on different data structures in order to investigate different available XAI methods with distinct capabilities for both data structures in future work (see Section 5). As illustrated in Figure 2, the first investigated model represents a one-dimensional CNN trained on multivariate ECG time series data. In contrast, the second model is based on the GoogLeNet [26] and is trained on images showing the ECG signals. Both approaches obtain the data from the PTB-XL benchmarking dataset [28]. Before evaluating and comparing both disparate approaches, we describe the PTB-XL dataset, the applied data preprocessing, and both mentioned models in the following sections.

3.1 Data Analysis & Preprocessing

The PTB-XL benchmarking dataset comprises the largest publicly available collection of ECG data, consisting of 21799 12-lead ECG records of ten seconds from a diverse cohort of 18869 patients [28]. The ECGs are accompanied by diagnostic labels provided by two cardiologists. Each ECG is labelled with possibly multiple superclasses, including myocardial infarction (MI), normal ECG (NORM), ST/T change (STTC), conduction disturbance (CD) and hypertrophy (HYP). There is also given a likelihood for each label being correct. We only make use of those ECG records that are labelled with MI or NORM (and possibly other superclasses) with a probability of 80% or more. We call this first preprocessing step filtering. This step is followed by an aggregation of labels into NORM

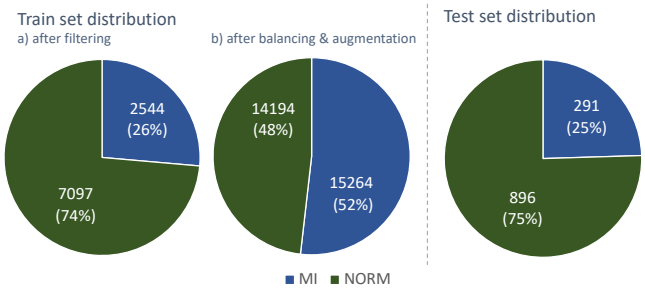

Figure 3. Class distribution of train and test sets after different steps of preprocessing.

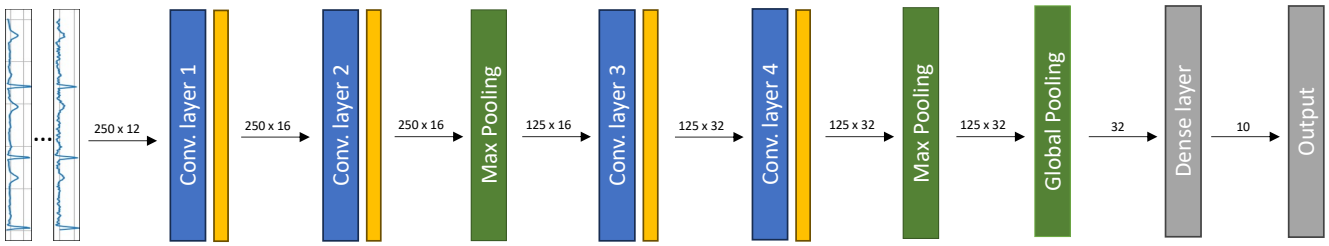

Figure 4. Architecture of the 1D CNN showing convolutional (blue), dropout (yellow), pooling (green) and dense (grey) layers.

and MI as we aim for a model that is capable to distinguish between these two classes. In the next step, we split the dataset into training (80%), validation (10%) and test (10%) as suggested and defined by Wagner et al. [28]. As illustrated in Figure 3, the resulting classes MI and NORM are unbalanced (1 : 3) in the training set. To address this imbalance issue and allow the models to learn more of the rare but targeted MI cases during training, we make use of a sliding window approach similar to that proposed by Strodthoff et al. [25]. As can be inferred from Figure 3, we obtain an almost balanced training set with ECGs of 2.5 seconds length. To be more precise, we oversample the minority class by selecting three subsequences of length 2.5s from each ECG with MI for each subsequence that we take from an ECG classified as NORM. As we additionally want to augment the whole training set in order to increase the learning capability of the ML models, we use two distinct subsequences from each ECG that is classified as NORM and six overlapping subsequences that are evenly distributed over the whole ECG sequence from each MI sample. The resulting data distribution and number of samples per class before and after preprocessing of the training set is depicted in Figure 3. As we can also see in this figure, we do not apply any balancing or augmentation on the test set as it has to represent the original data distribution to be expected in operation environments in order to derive reliable conclusions about the generalizability of the models. However, we also selected (random) subsequences of 2.5s to match the expected input dimension implied by the windowing approach. In this process, we ensured that no samples from individual patients were mixed in the individual data splits to avoid inter-patient learning of the models.

3.2 Approach 1: Using CNNs to analyse ECG signals

In our first approach, we use the preprocessed ECG signals as input to train a CNN with one-dimensional convolutional layers operating on the original multivariate time series data. This CNN is an adapted version of the one proposed by Hammad et al. [9], which they used as a tool for extracting deep features of ECG signals. The architecture of our adapted CNN version is illustrated in Figure 4. It consists of four convolutional layers, two max pooling layers, one global average pooling layer that is followed by a dense layer and a probabilistic output given by the sigmoid function. For all other layers, we use ReLU activations. We incorporated dropout layers after each of the convolutional layers as a regularization technique to reduce overfitting [23]. Additionally, we applied early stopping in order to avoid overfitting. To be more precise, we monitored the loss on the validation set and ended training the model after 30 iterations of no decrease. In order to obtain a model of high performance, we tuned the hyperparameters kernel sizes, dropout rates, batch size and learning rate for the Adam optimizer used during training. We thereby applied grid search resulting in kernel sizes of 7, 5, 5 and 5 for the four convolutional layers, dropout rates of 0.5, 0.4, 0.3 and 0.2 for the four dropout layers, a batch size of 64 and a learning rate of 0.001.

3.3 Approach 2: Using GoogLeNet to analyse ECG images

In this approach, we use images of ECG signals instead of the raw ECG signals, and use the GoogLeNet [26] for the classification task. GoogLeNet is a CNN with 22 layers using repetitive components of multi-sized filters at the same level. As an input to this model, we use images generated from the preprocessed data as described in Section 3.1.Thus, we ensure similar training data for both approaches. As GoogLeNet requires input images of size $(224, 224, 3)$, we resize the images before feeding them into the network. In addition, as the classification task is different from the original GoogLeNet classification, we modified it to be retrained for the MI detection task. We removed the last 'loss-3-classifier' layer and added a 'fully-connected layer', a 'softmax layer', and a 'classification layer' specific to our task. The rest of the network remained unchanged. We used a stochastic gradient descent optimizer with a momentum of 0.9 for optimization, with an initial learning rate of $3 \cdot 10^{-4}$ and batch size of 10. To avoid overfitting, we used the validation set as in approach 1.

4 Results and Discussion

We evaluated our two models on the recommended PTB-XL test set consisting of 291 MI and 896 NORM samples (see Figure 3). The results for both proposed models are shown in Table 1. They are based on a random 2.5s long subsequence for each of the ECGs from the test set. Figure 5 shows that the choice of the 2.5s subsequence does not have a great impact on the performance metrics accuracy, recall and precision.

Precisely, it depicts statistics of these metrics for both the models based on the test set. We used sliding windows for this analysis with a stride of 10 ms for the 1D CNN, and 500 ms for the GoogLeNet due to the computational complexity for generating images.

As we can deduce from Table 1, the accuracy of the 1D CNN model is with 96.21% slightly higher than that of the GoogLeNet, which manifests an accuracy of 95.53%. However, this metric should be used with caution as the test set is imbalanced. Therefore, our main focus is on the two performance metrics precision and recall. On the one hand, we aim for models with high precision identifying what proportion of ECGs classified as MI was actually correct. More specifically, we want the model not causing too many false alarms when incorporated in the eCall system. On the other hand, we aim for models with high recall, meaning the proportion of actual MIs being identified as such should be high. This is of high importance for our application as every non-identified MI endangers somebody's life. Whereas the precision of the 1D CNN with 91.55% is higher than that of GoogLeNet with 88.14%, the recall of the GoogLeNet is higher than that of the 1D CNN with 94.50% and 93.13%, respectively. For the sake of completeness, we also included other performance metrics in Table 1. Specificity, for example, is often used together with recall (sensitivity) in the clinical context. Regarding the performance metrics, we conclude that the 1D CNN manifests a slightly better trade-off between precision and recall than the GoogLeNet for the eCall use case. However, performance is not the only aspect that we take into account for our application. We additionally aim for an accountable model providing reasoning behind its decisions. An outlook on this aspect is given in Section 5.

4.1 Selecting different ECG windows at inference time

As described above, our proposed models expect ECG inputs of 2.5s length. However, ECGs with longer sequences are often available in real-world scenarios. Therefore, longer

Table 1. Performance comparison of the proposed models

Metric	1D CNN (Approach 1)	GoogLeNet (Approach 2)
Accuracy	96.21%	95.53%
Precision	91.55%	88.14%
Recall	93.13%	94.50%
Specificity	97.21%	95.87%
AUROC	98.91%	99.09%
TP	271	275
FP	25	37
TN	871	859
FN	20	16

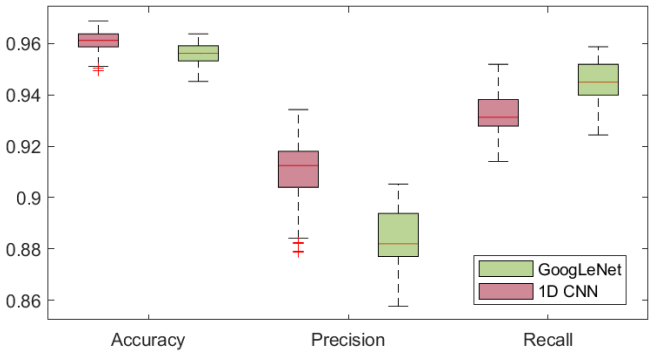

Figure 5. Accuracy, precision and recall for 1D CNN and GoogLeNet for sliding windows of 2.5 seconds over 10 seconds of ECG data. The boxes represent the 25th to 75th percentile of the samples, red lines represent the median, extended black lines represent the minimum and maximum, and red crosses represent the outliers.

Table 2. Comparison of test performance with different fragmentation approaches for the proposed 1D CNN model (A1) and the GoogLeNet-based model (A2).

Metric	Random window		Average windows		Max window	
	A1	A2	A1	A2	A1	A2
Accuracy (%)	96.08	95.59	96.54	95.87	95.28	93.93
Precision (%)	91.03	88.37	92.23	89.03	85.71	81.19
Recall (%)	93.23	94.46	93.81	94.84	96.91	97.93
TP	271	275	273	276	282	285
FP	27	36	23	34	47	66
TN	869	860	873	862	849	830
FN	20	16	18	15	9	6

ECGs could be fragmented into consecutive 2.5s windows, on which the MI detection models are applied continuously. For the presented emergency detection use case, this concept raises the question of when a fragmented ECG should be classified as MI and an emergency call should be triggered. To investigate and clarify this question, we examine and compare three fragmentation approaches on the PTB-XL test set. In the first approach, only a single random 2.5s window is selected from the 10s original ECG, and then used for MI detection. In the second approach, the ECG data is split into four consecutive 2.5s windows, and the final classification is based on averaging the prediction results of the individual windows. The third approach is similar to the second one, except that the ECG is classified as MI and an alarm is issued if at least one window is classified as MI.

Table 2 lists the results for the three discussed fragmentation approaches for both proposed models. For the first approach in which a random subwindow is chosen, we provide

the average metrics over all possible selections of windows as depicted in Figure 5. As we can deduce from Table 2, the performance metrics accuracy, precision, recall, TP, FP, TN and FN are all slightly better for the averaging approach than for the random window approach. However, the models will be integrated in a time-critical emergency system and evaluating four sequences instead of one would increase the classification time. The third approach (max window), in which the final classification is also based on analysing four windows, shows a possibility on how to increase the recall with the cost of decreasing the precision. The accuracy for this third approach is slightly smaller than that of the other two approaches. In conclusion, there is a trade-off between time and performance as well as between recall and precision that both needs to be taken into account when integrating the models into the eCall system.

5 Towards accountable ML-based eCalls

For our safety-critical application of automated emergency detection, the end-to-end accountability of the entire AI-based system is of utmost importance. Precisely, we want the ML models to be able to explain their decisions when predicting emergencies in a way that they are understandable by the emergency call takers. In our studied exemplary use case of MI detection, we want to provide visual explanations highlighting the relevant segments in the ECG signal that are indicative for a classification as MI. Therefore, we focus on those XAI methods that return an importance score for each point in time and each of the 12 leads or for each pixel, depending on the approach. This importance score represents the relevance of each input feature for the model's decision for detecting MI, allowing to present explanations as heatmaps superimposing the ECG being explained. In the following section, we aim to present the initial results and findings of investigating the application of XAI methods to the proposed models for realizing the explanations described above. A complete and systematic study of the models' accountability is subject to the current research activities of the authors and will be presented in future work.

5.1 Utilizing XAI methods for time series ECGs

In recent years, various explainable artificial intelligence methods [21] have been proposed that can be applied to different types of ML models and different types of data. The PTB-XL dataset that we use is a set of multivariate time series. However, most research and available XAI methods have focused on text, image, and tabular data [21]. Although univariate time series data can be interpreted as tabular data and, therefore, XAI methods for tabular data can be transferred to them, this is not straightforward in the case of multivariate time series data. A dataset of multivariate time series has three dimensions: the number of time series data, the number of time steps, and the value at each time step. In contrast to that, tabular data only has two dimensions. There is the possibility to reduce one dimension of the multivariate time series dataset by flattening each time series but this comes with the cost of losing explicit time dependencies of features, which is why we decided against this approach. At this point, we want to remind the reader that this XAI limitation of multivariate time series classifiers is the reason for presenting two models working on different data structures.

For the proposed 1D CNN model discussed in Section 3.2, we encountered LRP [2] and SHAP [13] to be applicable to explain the model's decisions on the multivariate time series. Figure 6 illustrates exemplary explanations for an ECG of one patient for which the 1D CNN model correctly predicted an MI with a classification score of 99.99%. We scaled each of the importance scores to be in the range $[-1, 1]$ by dividing with the highest absolute value of all importance scores per explanation in order to have comparable scores across XAI methods while keeping the sign. The sign is important as negative values in LRP denote negative evidence for a class [2], i.e. evidence for the respective ECG being normal in our case. The two explanations given in Figure 6 highlight segments encountered as most relevant regarding the 1D CNN model's decision for predicting MI in red colour. As illustrated, there seems to exist some consistency of red highlighted segments across heart beats and some of the leads. Additionally, some of the 12 leads have been highlighted more, suggesting to be more relevant for detecting MI.

5.2 Utilizing XAI methods for image ECGs

For the proposed GoogLeNet-based model presented in Section 3.3, we use ECG images as input to perform the MI detection. Thus, this image classifier can potentially exploit the wide range of available XAI methods for models working on image data [21]. For the presented model implemented in Matlab, we found Grad-CAM [22] and LIME [19] as the

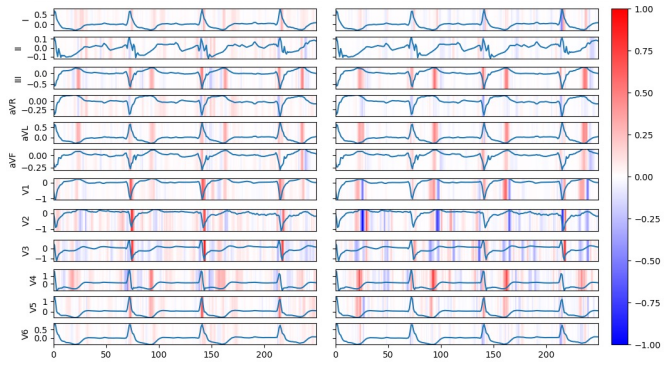

Figure 6. Explanations from SHAP (left) and LRP-epsilon (right) for the output of the 1D CNN (approach 1) on the same ECG. The horizontal axis represents time in 10ms and the vertical axis represents voltage in mV. Red represents positive whereas blue represents negative relevance for MI.

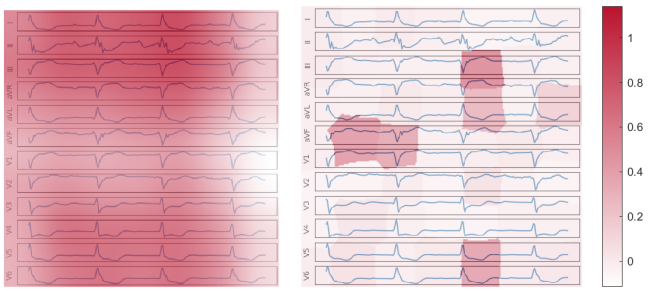

Figure 7. Explanations from Grad-CAM (left) and LIME (right) for the output of GoogLeNet (approach 2). The red colour represents the most relevant parts of the ECG image.

initially most promising approaches. Figure 7 shows the explanations generated by the two XAI methods for the same patient as used in Figure 6. The GoogLeNet prediction for this image was MI with a confidence score of 99.61%. The relevance values in the image are scaled as in the first approach. Since Grad-CAM and LIME had only positive scores, they appear to be in the range $[0, 1]$, where 0 indicates no influence, while 1 indicates most relevance towards classifying MI. As observed in Figure 7, Grad-CAM highlights a larger area and is not able to highlight specific parts of the image. For various other ECG images, Grad-CAM was able to provide more specific areas than in this example. However, LIME highlights specific leads and segments of the ECG image, providing more specific parts that may be relevant.

5.3 Outlook: Evaluate model accountability

When comparing the results illustrated in Figures 6 and 7, we can conclude that the two explanations for the 1D CNN model's decision on detecting MI are fine-grained, whereas those for the GoogLeNet model's decision are more coarse-grained. The explanations given for the 1D CNN mark individual ECG segments as relevant. On the contrary, the explanations of the GoogLeNet highlight broader segments or whole leads. However, there also seems to exist some consistency of highlighted segments across these XAI methods.

So far we have shown the capability to provide explanations for each of the models, but a thorough quantitative and qualitative analysis and comparison of both approaches and the generated explanations is yet open and currently conducted by the authors. However, the findings presented in this section already highlight several research questions that remain open and will be investigated in future work: 1) Which of the XAI methods applicable to one model provides the best explanations? 2) Which of the two proposed models is more accountable, i.e. provides better explanations? 3) Which of the models and methods generates stable and consistent explanations? While the first two research questions describe a qualitative analysis for which domain experts need to be included, the third question corresponds to a quantitative analysis for which suitable metrics to measure stability and consistency need to be defined.

6 Conclusion

In this paper, we proposed two disparate ML models capable of identifying indications for MIs from ECG sensor data in the context of IoT-enabled emergency communication systems. While both models reliably detect MIs, we found that the presented 1D CNN operating on time series data shows a better trade-off between precision and recall for the discussed eCall use case than the GoogLeNet-based model working on plotted ECGs. Furthermore, we analyzed three ECG fragmentation approaches for consecutively applying the MI detection models in real-world scenarios. Our findings indicate that no approach outperforms the others. Instead, a trade-off between classification time and performance needs to be found when integrating the models into safety-critical systems. Finally, we presented the initial results of applying XAI methods to the proposed models for providing explanations that elucidate their decision-making. These preliminary experiments suggest that the CNN model's explanations for detecting MIs are more fine-grained in comparison to the GoogLeNet-based model. In future work, we aim to investigate the last aspect in more detail by analysing how to approach accountability for the proposed models by improving their explainability. Precisely, we want to systematically evaluate the application of XAI methods to the proposed models by performing quantitative and qualitative analysis of the provided explanations as described in Section 5.

Acknowledgments

This research is supported by SPATIAL project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101021808.

References

[1] Atul Anand, Tushar Kadian, Manu Kumar Shetty, and Anubha Gupta. 2022. Explainable AI decision model for ECG data of cardiac disorders. *Biomedical Signal Processing and Control* 75 (2022), 103584. https://doi.org/10.1016/j.bspc.2022.103584

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[3] Prachi Bagave, Marcus Westberg, Roel Dobbe, Marijn Janssen, and Aaron Yi Ding. 2022. Accountable AI for Healthcare IoT Systems. In *2022 IEEE 4th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*. 20–28. https://doi.org/10.1109/TPS-ISA56441.2022.00013

[4] Ramon Barakat, Faruk Catal, Nikolay Tcholtchev, and Yacine Rebahi. 2020. TTCN-3 based NG112 Test System and Playground for Emergency Communication. In *2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*. 492–497. https://doi.org/10.1109/QRS-C51114.2020.00088

[5] X. Chen, W. Guo, L. Zhao, W. Huang, L. Wang, A. Sun, L. Li, and F. Mo. 2021. Acute Myocardial Infarction Detection Using Deep Learning-Enabled Electrocardiograms. *Front Cardiovasc Med* 8 (2021), 654515. https://doi.org/10.3389/fcvm.2021.654515

[6] Eurostat. 2023. *Population structure and ageing.* https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_structure_and_ageing

[7] R. Fang, C. C. Lu, C. T. Chuang, and W. H. Chang. 2022. A visually interpretable detection method combines 3-D ECG with a multi-VGG neural network for myocardial infarction identification. *Comput Methods Programs Biomed* 219 (2022), 106762. https://doi.org/10.1016/j.cmpb.2022.106762

[8] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation* 101, 23 (2000), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

[9] Mohamed Hammad, Samia Allaoua Chelloug, Reem Alkanhel, Allam Jaya Prakash, Ammar Muthanna, Ibrahim A Elgendy, and Paweł Pławiak. 2022. Automated detection of myocardial infarction and heart conduction disorders based on feature selection and a deep learning model. *Sensors* 22, 17 (2022), 6503.

[10] C. Han and L. Shi. 2020. ML-ResNet: A novel network to detect and locate myocardial infarction using 12 leads ECG. *Comput Methods Programs Biomed* 185 (2020), 105138. https://doi.org/10.1016/j.cmpb.2019.105138

[11] V. Jahmunah, E. Y. K. Ng, R. S. Tan, S. L. Oh, and U. R. Acharya. 2022. Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals. *Comput Biol Med* 146 (2022), 105550. https://doi.org/10.1016/j.compbiomed.2022.105550

[12] Budankailu Kumar Subudhi, Faruk Catal, Nikolay Tcholtchev, Kin Tsun Chiu, Yacine Rebahi, Michell Boerger, and Philipp Lämmel. 2019. Performance Testing for VoIP Emergency Services: a Case Study of the EMYNOS Platform and a Reflection on potential Blockchain Utilisation for NG112 Emergency Communication. *Journal of Ubiquitous Systems and Pervasive Networks* 12, 1 (Nov. 2019), 01–08. https://doi.org/10.5383/JUSPN.12.01.001

[13] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[14] Xin Ma, Xingwen Fu, Yiqi Sun, Nan Wang, and Yang Gao. 2023. Application of Convolutional Dendrite Net for Detection of Myocardial Infarction Using ECG Signals. *IEEE Sensors Journal* 23, 1 (2023), 460–469. https://doi.org/10.1109/jsen.2022.3221779

[15] H. Makimoto, M. Hockmann, T. Lin, D. Glockner, S. Gerguri, L. Clasen, J. Schmidt, A. Assadi-Schmidt, A. Bejinariu, P. Muller, S. Angendohr, M. Babady, C. Brinkmeyer, A. Makimoto, and M. Kelm. 2020. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. *Sci Rep* 10, 1 (2020), 8445. https://doi.org/10.1038/s41598-020-65105-x

[16] H. Martin, U. Morar, W. Izquierdo, M. Cabrerizo, A. Cabrera, and M. Adjouadi. 2021. Real-time frequency-independent single-Lead and single-beat myocardial infarction detection. *Artif Intell Med* 121 (2021), 102179. https://doi.org/10.1016/j.artmed.2021.102179

[17] Hari Mohan Rai and Kalyan Chatterjee. 2021. Hybrid CNN-LSTM deep learning model and ensemble technique for automatic detection of myocardial infarction using big ECG data. *Applied Intelligence* 52, 5 (2021), 5366–5384. https://doi.org/10.1007/s10489-021-02696-6

[18] Yacine Rebahi, Kin Tsun Chiu, Nikolay Tcholtchev, Simon Hohberg, Evangelos Pallis, and Evangelos Markakis. 2018. Towards a next generation 112 testbed: The EMYNOS ESInet. *International Journal of Critical Infrastructure Protection* 22 (Sept. 2018), 39–50. https://doi.org/10.1016/j.ijcip.2018.05.001

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778 event-place: San Francisco, California, USA.

[20] Joana Schmidt. 2019. Weltweit häufigste Todesursache: Überholt Krebs kardiovaskuläre Erkrankungen? *CardioVasc* 19, 5 (Oct. 2019), 11–11. https://doi.org/10.1007/s15027-019-1617-y

[21] Gesina Schwalbe and Bettina Finzel. 2023. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery* (Jan. 2023). https://doi.org/10.1007/s10618-022-00867-8

[22] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[24] Nils Strodthoff and Claas Strodthoff. [n. d.]. Detecting and interpreting myocardial infarction using fully convolutional neural networks. ([n. d.]).

[25] Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. 2020. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics* 25, 5 (2020), 1519–1528. https://doi.org/10.1109/JBHI.2020.3022989

[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[27] R. Uchiyama, Y. Okada, R. Kakizaki, and S. Tomioka. 2022. End-to-End Convolutional Neural Network Model to Detect and Localize Myocardial Infarction Using 12-Lead ECG Images without Preprocessing. *Bioengineering (Basel)* 9, 9 (2022). https://doi.org/10.3390/bioengineering9090430

[28] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima Lunze, Wojciech Samek, and Tobias Schaeffter. 2020. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data* 7 (05 2020), 154. https://doi.org/10.1038/s41597-020-0495-6

[29] Jieshuo Zhang, Ming Liu, Peng Xiong, Haiman Du, Hong Zhang, Feng Lin, Zengguang Hou, and Xiuling Liu. 2021. A multi-dimensional association information analysis approach to automated detection and localization of myocardial infarction. *Engineering Applications of Artificial Intelligence* 97 (2021). https://doi.org/10.1016/j.engappai.2020.104092

[30] Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. 2020. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data* 7, 1 (Feb. 2020), 48. https://doi.org/10.1038/s41597-020-0386-x Number: 1 Publisher: Nature Publishing Group.