Exploring Interpretable AI Methods for ECG Data Classification

Jaya Ojha
Department of Computer Science
OsloMet – Oslo Metropolitan University
Oslo, Norway
s371136@oslomet.no

Hårek Haugerud
Department of Computer Science
OsloMet – Oslo Metropolitan University
Oslo, Norway
NordSTAR - Nordic Center for Sustainable and Trustworthy AI Research
Oslo, Norway
harek.haugerud@oslomet.no

Anis Yazidi
Department of Computer Science
OsloMet – Oslo Metropolitan University
Oslo, Norway
OsloMet Artificial Intelligence Lab
Oslo, Norway
NordSTAR - Nordic Center for Sustainable and Trustworthy AI Research
Oslo, Norway
anisy@oslomet.no

Pedro G. Lind
Department of Computer Science
OsloMet – Oslo Metropolitan University
Oslo, Norway
OsloMet Artificial Intelligence Lab
Oslo, Norway
NordSTAR - Nordic Center for Sustainable and Trustworthy AI Research
Oslo, Norway
Simula Research Laboratory, Numerical Analysis and Scientific Computing
Oslo, Norway
pedro.lind@oslomet.no

ABSTRACT

We address ECG data classification, using methods from explainable artificial intelligence (XAI). In particular, we focus on the extended performance of the ST-CNN-5 model compared to established models. The model showcases slight improvement in accuracy suggesting the potential of this new model to provide more reliable predictions compared to other models. However, lower values of the specificity and area-under-curve metrics highlight the need to thoroughly evaluate the strengths and weaknesses of the extended model compared to other models. For the interpretability analysis, we use Shapley Additive Explanations (SHAP), Gradient-weighted Class Activation Mapping (GradCAM), and Local Interpretable Model-agnostic Explanations (LIME) methods. In particular, we show that the new model exhibits improved explainability in its GradCAM explanations compared to the former model. SHAP effectively highlights crucial ECG features, better than GradCAM and LIME. The latter methods exhibit inferior performance, particularly in capturing nuanced patterns associated with certain cardiac conditions. By using distinctive methods in the interpretability analysis, we provide a systematic discussion about which ECG features are better - or worse - uncovered by each method.

ICDAR '24, June 10–14, 2024, Phuket, Thailand
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0549-6/24/06
https://doi.org/10.1145/3643488.3660294

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**.

KEYWORDS

Artificial Intelligence, Deep Learning, CNN, Explainable AI, SHAP, GradCAM, LIME

ACM Reference Format:
Jaya Ojha, Hårek Haugerud, Anis Yazidi, and Pedro G. Lind. 2024. Exploring Interpretable AI Methods for ECG Data Classification. In *The Fifth Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR '24), June 10–14, 2024, Phuket, Thailand.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3643488.3660294

1 INTRODUCTION

Electrocardiogram (ECG) has been used in medicine for over a century and remains an essential diagnostic tool for heart conditions [29]. It can offer insightful information about both cardiac and non-cardiac health and diseases, including arrhythmias, ischemia, hypertrophy, and electrolyte disturbances. However, the ECG interpretation requires extensive knowledge and experience and is susceptible to errors and biases of expert (human) analysis [27]. Therefore, a correct interpretation of the ECG is essential, which requires healthcare professional expertise able to recognize subtle changes in the electrical activity of the heart [3]. However, medical professionals typically detect heart arrhythmias by visually inspecting the ECG wave patterns and shapes, which can be time-consuming and resource-expensive. Moreover, such empirical ECG analysis is prone to misinterpretations, particularly when there is a lot of data to examine and decisions are expected to be taken in short time periods, raising risks of misdiagnosis of a fatal heart condition [34].

Artificial Intelligence (AI) has emerged as a promising tool in ECG analysis [27], offering the potential to improve the accuracy and efficiency of diagnoses and treatment plans. AI systems can recognize patterns of illness and predict accurate diagnoses. These algorithms have in some cases outperformed the trained eye of clinical experts in identifying high-risk patients [19]. This is due to the ability of AI to differentiate between heart diseases, based on the identification of patterns, which are very similar and have small discrepancies hidden from the human eye [1].

However, the high performance of AI models comes with a cost: the lack of transparency. The high complexity of AI models - particularly those that follow the so-called deep learning (DL) paradigm [10]- obscures the rationale behind the decisions retrieved as an outcome of an algorithm [8]. The opacity of such algorithms affects the trust of physicians accountable for the medical decisions, and therefore undermines the usability of such AI solutions in practice.

Improving the interpretability of deep learning classifiers in the medical domain presents challenges owing to the inherent uncertainty, class imbalance, data heterogeneity, and noise present in medical datasets [7]. Since it is improbable that AI will substitute the full role of cardiovascular professionals, AI must be trustful from the expert perspective and to be trustful for a healthcare professional, it must provide interpretable outcomes [21].

In this paper, we analyze ECG data using deep learning algorithms to detect heart diseases and assess their interpretability, using a panoply of different methods in Explainable Artificial Intelligence (XAI) whose output is then discussed and compared.

2 RELATED WORK

Deep learning methods have transformed cardiac disease classification [24], offering advanced capabilities to analyze complex medical data. We explore some DL methodologies for classifying cardiac diseases, followed by an examination of XAI techniques used to interpret DL model decisions. Through reviewing related works, we aim to highlight advancements and challenges in leveraging DL and XAI for enhanced cardiac diagnosis.

2.1 DL methods to classify cardiac diseases

Many studies focus on the utilization of DL techniques for analyzing ECG data. Somani et al. [28] conducted a review of 31 research papers and found that Convolutional Neural Networks (CNNs) were extensively used for ECG analysis due to their effectiveness in identifying patterns in large healthcare datasets. Their review highlighted the groundbreaking performance of deep learning models in uncovering hidden associations within vast datasets, showcasing the promising potential of these models in revolutionizing ECG analysis. Additionally, Khurshid et al. [18] developed a deep learning-based model for predicting the risk of atrial fibrillation (AF), based on 12-lead ECG data. Their study demonstrated the reliability and validity of deep learning models in estimating AF risk across diverse populations, suggesting their potential as valuable tools in quantifying future AF risk.

Kashou et al. [17] developed an artificial intelligence-enabled ECG (AI-ECG) algorithm capable of analyzing 12-lead ECGs as accurately as cardiologists. Their study showcased the potential of AI-ECG systems in improving ECG interpretation, reducing medical errors, and enhancing clinical workflows. Additionally, another study [5] utilized artificial intelligence-enhanced ECG methods to identify patients with AF by developing a CNN capable of detecting AF signatures present during sinus rhythm. These findings underlined the potential of AI-enabled ECG analysis in accurately diagnosing AF, thereby aiding in better patient management. An article [27] reviewed the application of artificial intelligence in enhancing electrocardiography (AI-ECG) for cardiovascular disease management, highlighting its role as a potent non-invasive biomarker for cardiovascular diseases.

While such works provide significant evidence of the technical performance of DL algorithms to classify ECGs, the level of "acceptance" and their reliability from the perspective of human experts does not necessarily match this level of performance, due to the need for interpretability.

2.2 XAI methods

XAI is a rapidly growing field of research, driven by concerns about the safety and ethics of using high-performance "black-box" AI models. The need for explainability arises from public apprehension about AI's potential consequences. Researchers view explainability as a way to enhance trust and acceptance of AI systems. Qualitative and quantitative studies on human-AI interactions seek to validate and/or assess this trust enhancement through explainability, among other aspects [37] [31].

Benchmark methods in XAI include the SHapley Additive exPlanations (SHAP), by Anand et al. [4], which enhances the interpretability of deep neural network models developed for ECG analysis. This transparency not only fosters trust and confidence among healthcare professionals, particularly cardiologists but also enables them to understand the model's decision-making process and validate its outputs against their expertise. The study emphasizes the potential impact of interpretable AI models in low- and middle-income countries, where healthcare infrastructure is often strained, by improving the efficiency of healthcare services through the integration of interpretable AI models into clinical workflows.

Another important method in XAI, introduced by Hicks et al. [12], is the ECGradCAM, a method to enhance the interpretability of deep learning-based ECG analysis, which generates attention maps to explain neural network decisions and reveals insights into their operation. The study highlights the importance of interpretable machine learning models in medicine and advocates for their thorough evaluation and adoption across medical domains beyond ECG analysis. Furthermore, Agrawal et al. [2] proposed an XAI solution to enhance the explainability of heartbeat classification, aiming to explain the rationale behind classification decisions using various model-agnostic methods. Their study stressed the importance of interpretability in medical AI applications and suggested avenues for refining and extending the proposed methodology for real-world implementation in clinical settings.

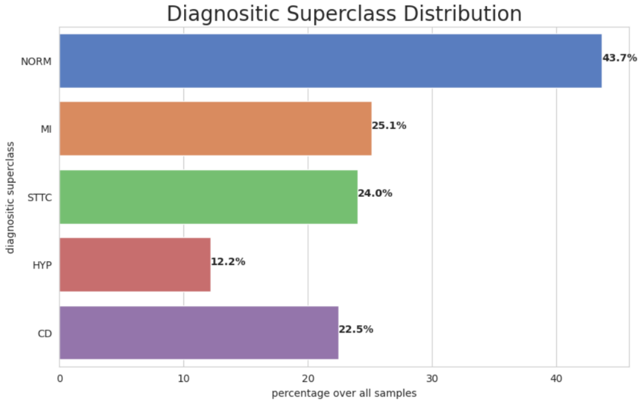

Figure 1: Superclass distribution of the data set that shows "Normal" ECG records have the highest percentage, followed by "MI", and then "STTC". "CD" and "HYP" contain the fewest records. This distribution suggests that the model may have greater accuracy in diagnosing "Normal" cases, given the larger number of available records for this category, compared to "HYP".

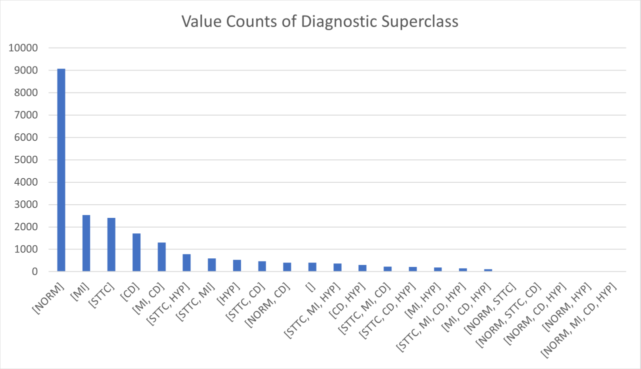

Figure 2: Plot of value counts of Superclass Diagnosis. A single ECG record may be associated with as many as four different superclasses. This shows the multilabeled nature of the dataset, where each ECG record can be categorized under multiple superclasses simultaneously.

3 MATERIALS AND METHODS

3.1 The Data set

In this work, we use an openly available dataset provided by the Physikalisch-Technische Bundesanstalt (PTB) [33]. PTB built a large database of ECG records, called the PTB-XL ECG data set. The database comprises a total of 21,801 clinical 12-lead ECG records of 10 seconds in length from 18,869 patients. This diverse patient population is almost evenly split between genders, with 52% being male and 48% female. The data set offers a wide age spectrum, ranging from newborns to individuals aged 95, with a median age of 62 and an interquartile range spanning 22 years.

Each of the records falls into one or more of five distinct superclass categories, forming the basis for multilabel classification. The five diagnostic superclass categories of this data set are "NORM" for normal ECG readings, "MI" representing myocardial infarction, "STTC" denoting ST/T changes in ECGs, "CD" indicating conduction disturbances, and "HYP" signifying hypertrophy. The value counts and distribution of these various categories are visually represented in Figures 1 and 2.

Metadata and waveforms were transferred to open data formats for simple processing by common applications. The waveform files are kept in WaveForm DataBase (WFDB) format, which has a sampling rate of 500 Hz and a downsampled version of the waveform data at a sampling frequency of 100Hz, a resolution of 1 LSB, and 16-bit precision [33].

3.2 Implementation of DL architectures

The datasets described above were trained and evaluated with the deep learning architectures described in this section.

3.2.1 *Spatio-Temporal CNN Model (ST-CNN).*

We recreated the ST-CNN-5 Model based on the paper by Anand et al. [4]. The ’5’ signifies the inclusion of five temporal layers within its architectural design. This model comprises of convolutional layers for feature extraction from ECG data with 12-time steps and 1000 features, integrating skip connections for gradient flow. It splits into five temporal and one spatial analysis paths, with convolutional layers capturing temporal dependencies and global average pooling aggregating spatial information. Following this, two fully connected layers perform high-level feature abstraction, while the implementation of a L2-regularization and dropout criteria mitigate overfitting. The model concludes with a dense layer applying the sigmoidal activation function for multi-label classification, yielding a probability distribution over classes.

In this work, we extend this ST-CNN-5 model, hereafter referred to as "ST-CNN-5 new" going forward, by incorporating the following:

(1) Skip connections: Added skip connections before applying the ReLU activation function, allowing direct influence on subsequent layer outputs, thus addressing vanishing gradients and enhancing performance.
(2) EarlyStopping callback: Stops training if validation loss does not improve for a specified number of epochs, preventing overfitting.
(3) ReduceLROnPlateau callback: Dynamically adjusts learning rate during training for improved convergence.
(4) An increased dropout rate: Raised dropout rate to 0.3 to mitigate overfitting.

The model was configured using the Adam optimizer along with the binary cross-entropy loss function. Additionally, callback functions were utilized to optimize the model’s performance throughout the training process. The sketch illustrating the architecture of this extended model can be seen in Figure 3.

3.2.2 *Replication of state-of-the-art models.*

For a systematic comparison, we also consider benchmark models from the paper in Ref. [30], which fall into two categories: CNNs and recurrent neural networks (RNNs), both operating on the raw ECG signal. CNNs are further categorized into standard feed-forward architectures, resnet-based architectures, and inception-based architectures.

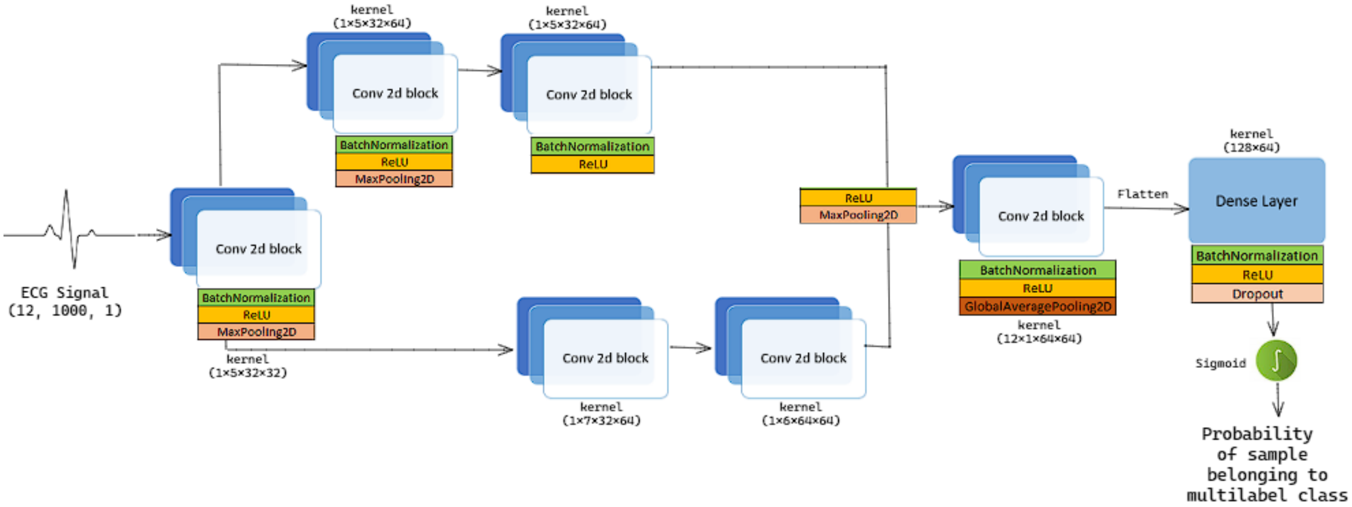

Figure 3: Block Diagram of "ST-CNN-5 new" Architecture: The model features five convolutional blocks dedicated to capturing temporal information, followed by a singular block focused on spatial analysis. Following this, the data flows through two fully connected layers incorporating a dropout rate of 0.3 for regularization. The final output is processed through a sigmoid activation function to derive the probabilities.

Authors from [30] mention that standard feed-forward architectures include fully convolutional networks [35] (fcn_wang) and resnet-based architectures [36] (resnet1d_wang) have demonstrated success in various large-scale studies. They also proposed one-dimensional adaptations and evaluated xresnets [11] (xresnet1dxxx). As a final convolutional architecture, they adapted InceptionTime [16](inception1d) architecture to the time series domain.

These implementations involved the use of a concat-pooling layer for pooling, aggregating the results of global average pooling and max pooling along the feature dimension. For resnets, they increased the kernel sizes to five. All convolutional models shared the same fully connected classification head with a single hidden layer comprising 128 hidden units, batch normalization, and dropout rates of 0.25 and 0.5 at the first and second fully connected layers, respectively [30].

For RNN, unidirectional and bidirectional LSTMs [13] (lstm, lstm_bidir) with two layers and 256 hidden units were considered. They aggregated the outputs using a concat-pooling layer. Binary cross-entropy optimization was employed which is suitable for multi-label classification problems. During training, 1-cycle learning rate scheduling and the Adam optimizer were utilized [30].

Their training process adopted the sliding window approach which is commonly used in time series classification. This involves training the classifier on random segments of fixed length from the full record, accommodating records of varying lengths, and effectively serving as data augmentation. During test time, they applied the test time augmentation dividing the record into segments with overlapping windows and aggregating model predictions using element-wise maximum. This aggregation significantly enhances overall performance compared to predictions on random sliding windows without aggregation. Unless specified otherwise, a fixed window size of 2.5 seconds was used [30].

All deep-learning models from this paper [30] were implemented using *PyTorch* [25], *fastai* library [14], and *Keras*.

3.3 Training, testing and performance metrics

The data set was split into training and testing subsets, where 10% was allocated for testing. Each model underwent training for 20 epochs, utilizing a batch size of 64 samples to optimize computational efficiency. Throughout the training process, the model's performance was continuously monitored using the specified callbacks to ensure optimal convergence and prevent overfitting. GPU was used as a hardware accelerator to handle computational demands efficiently. The ST-CNN models were implemented through the *Keras* API, which operates on the *Tensorflow* framework, offering a robust and streamlined platform for model development and experimentation.

The standard metrics to assess classification algorithm performance were used. Namely, from the number of true/false positives (TP/FP) and true/false negative (TN/FN), we compute:

- The ratio of correctly identified samples to all samples is called accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- The proportion of examples that are truly positive among all examples that we projected to be positive is called precision:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- Recall is the proportion of real positive instances to all positive cases:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

- Specificity refers to how well a test or model identifies true negatives among all the actual negative examples [15]:

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

Aditionally, we also use the Area Under The Curve (AUC) Receiver Operating Characteristics (ROC) curve as indicators of performance for classification at different threshold levels. These two metrics "mix" some of the metrics listed above. AUC is a measurement of separability, and ROC is a probability curve. It reveals how well the model can differentiate across classes. When AUC is high, the model is more accurate. The ROC curve plots the False Positive Rate (FPR) on the x-axis and the True Positive Rate (TPR) on the y-axis [23].

3.4 Interpretability

As discussed above, interpreting deep learning models, particularly CNNs, poses a significant challenge due to their complexity, often rendered as "black boxes". Understanding the features learned by these models is crucial for the so-called model transparency and trustworthiness. We use three widely recognized XAI methods: SHAP, Gradient-weighted Class Activation Mapping (GradCAM), and Local Interpretable Model-agnostic Explanations (LIME). These methods have been identified as among the most popular XAI techniques in healthcare, according to a systematic review spanning the last decade (2011–2022) [20].

We use the integrated gradients SHAP which calculates feature importance in deep neural networks by attributing contributions of each feature to the model's output [22].

We also use the GradCAM, which operates by constructing a model that maps input data to activations within the final convolutional layer [32]. We leveraged the gradients derived from the predicted class with respect to these activations which computes a weighted activation map. Through subsequent application of global average pooling, rectified linear unit (ReLU), and normalization techniques, the heatmap is refined, representing the spatial importance of features within the input data. Finally, the heatmap is resized to match the original input size enabling overlaying onto the original data, enhancing interpretability through the visualization of regions of interest, distinguished by lighter alpha values.

Finally, we also consider the *RecurrentTabularExplainer* from LIME [26]. This method estimates class probabilities. The respective plot delineates the relevance of features within the context of model predictions, offering valuable insights into the underlying decision-making process.

4 RESULTS AND DISCUSSION

4.1 Assessing performance: comparative analysis

In our comparative analysis between the "ST-CNN-5 new" model and other established models from existing literature, the results indicate its superior overall predictive performance compared to the other models evaluated as seen in Table 1.

Our "ST-CNN-5 new" architecture shows better accuracy and precision than the other models. While the "ST-CNN-5 new" model achieves higher accuracy, its specificity and AUC metrics are moderate, compared to the other models. Therefore, the overall performance of our extended model may not consistently surpass other models in terms of true negative instances, which is also reflected in lower AUC.

Table 1: Comparative Performance Analysis: Proposed Model vs. Literature Benchmarks

Model	accuracy	precision	recall	specificity	AUC
ST-CNN-5 [4]	0.885	0.796	0.665	0.932	0.924
ST-CNN-5 new	**0.891**	**0.798**	**0.693**	**0.934**	**0.932**
inception1d [30]	0.886	0.737	0.781	0.896	0.934
xresnet1d101 [30]	0.881	0.72	0.796	0.885	0.931
lstm_bidir [30]	0.884	0.738	0.773	0.895	0.931
resnet1d_wang [30]	0.879	0.704	0.803	0.884	0.929
lstm [30]	0.88	0.734	0.765	0.892	0.926
fcn_wang [30]	0.88	0.735	0.754	0.893	0.925

Table 2: Multiclass Classification report of "ST-CNN-5 new" Model for each class

Classes	precision	recall	specificity
CD	0.82	0.67	0.96
HYP	0.74	0.47	0.98
MI	0.85	0.69	0.96
NORM	0.82	0.91	0.85
STTC	0.77	0.73	0.93

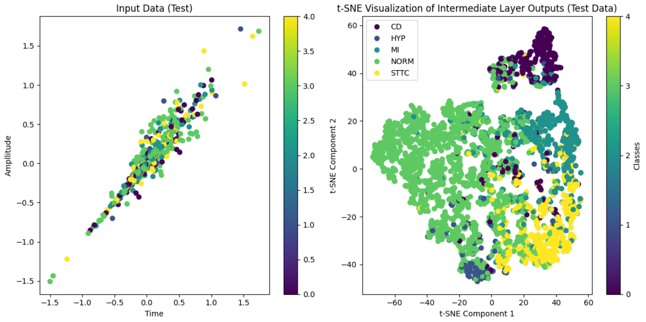

Figure 4: t-SNE plots illustrating the transformation of ECG data through the "ST-CNN-5 new" model. The first subplot visualizes the input ECG data after flattening, aligning its dimensions with the test dataset. In the second subplot, the t-SNE visualization showcases the output of the final convolutional layer, demonstrating the model's progressive learning of distinctive features across separated classes.

The main results concerning the multiclass classification of this new model are shown in Table 2. Both precision and specificity are considerably high, particularly the specificity, whereas recall shows a large value only for the control group, and it is particularly low for HYP classes. This indicates that except for the control group, the algorithm is better at detecting negative cases than positive

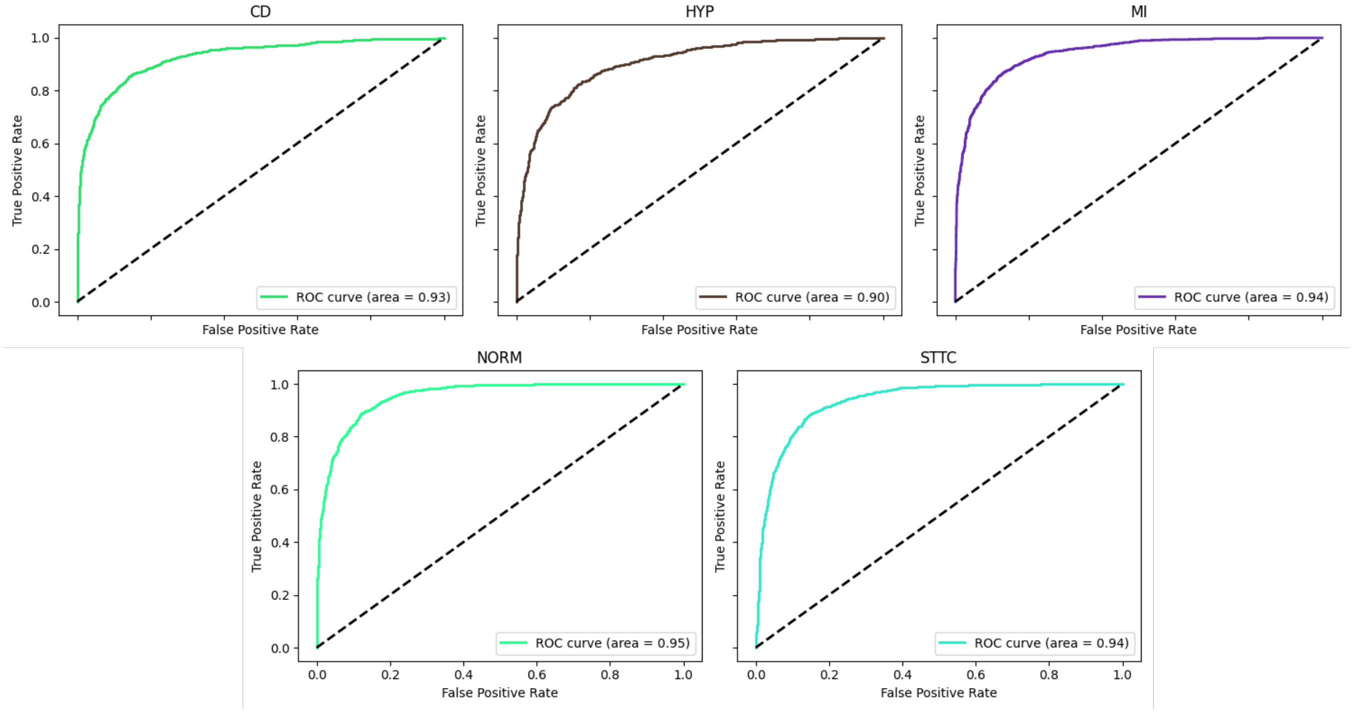

Figure 5: AUCROC curves of "ST-CNN-5 new" Model for each class

ones. The large values of both precision and specificity are related through the small number of false positives.

We have included t-distributed Stochastic Neighbor Embedding (t-SNE) plots in Figure 4 to showcase the ECG data's transformation process within the "ST-CNN-5 new" model. These plots are an unsupervised non-linear dimensionality reduction method for exploring and visualizing high-dimensional data [6]. The plots specifically focus on test records with a single class label, ensuring a clear representation. The visualization offers insight into the output of the final convolutional layer, illustrating how the "ST-CNN-5 new" model gradually learns distinctive features across separated classes. This shows that the model is good at finding important patterns in the ECG data, helping it classify different heart conditions accurately.

Looking into the AUCROC results of "ST-CNN-5 new" for each class, the model shows superior performance for the NORM class as seen in Figure 5. This emphasizes its capability to distinguish between normal and abnormal ECG recordings. Nevertheless, the model's performance is lower for conditions that involve more subtle ECG changes, such as HYP. This suggests that the model may struggle to detect these changes, especially in the presence of other factors that can affect ECG waveforms.

4.2 Interpretability of the new model

In this section, we evaluate the interpretability of the "ST-CNN-5 new" model, which showed overall higher performance than the other benchmarks.

The results are shown in Figure 6. For SHAP explanations, the visualization highlights ECG wave segments in red, indicating their importance with high SHAP values for the classification of the model. These segments are very important for the model's decision-making. Less significant features are shown in blue. In GradCAM and LIME heatmaps, darker areas signify the most influential contributions to the model's decision. These highlighted regions are specific to each lead, based on the data from that particular lead.

An intriguing observation revealed the improved interpretability of the "ST-CNN-5 new" model in its GradCAM explanations. The new model could explain its decisions better when compared to the old one.

In analyzing the characteristics of a normal ECG, sinus rhythm entails several distinct features, including the consistent positivity of the P wave and T wave, the presence of a net positive QRS complex, and a smooth transition from the ST segment to the T wave. It is represented by a positive deflection with a large, upright R wave in Lead II [9]. The SHAP explanation method effectively highlights the significance of the positive QRS complex and the smooth transitioning of the ST segment, showcasing their importance in ECG classification. Similarly, GradCAM demonstrates proficiency in recognizing these key features. However, the interpretation provided by the LIME explanation appears less robust in capturing these essential characteristics of normal sinus rhythm, suggesting potential limitations of LIME in accurately explaining ECG classifications. These observations can be seen in Row 1 of Figure 6.

The SHAP explanation pinpointed the deep S wave observed in Lead V1, a characteristic sign of Left Bundle Branch Block (LBBB) [9] as seen in Figure 6.2.b. when compared to the literature characteristic as seen in Figure 6.2.a.. This observation suggests a likely sinus rhythm with a prolonged PR interval, a characteristic often

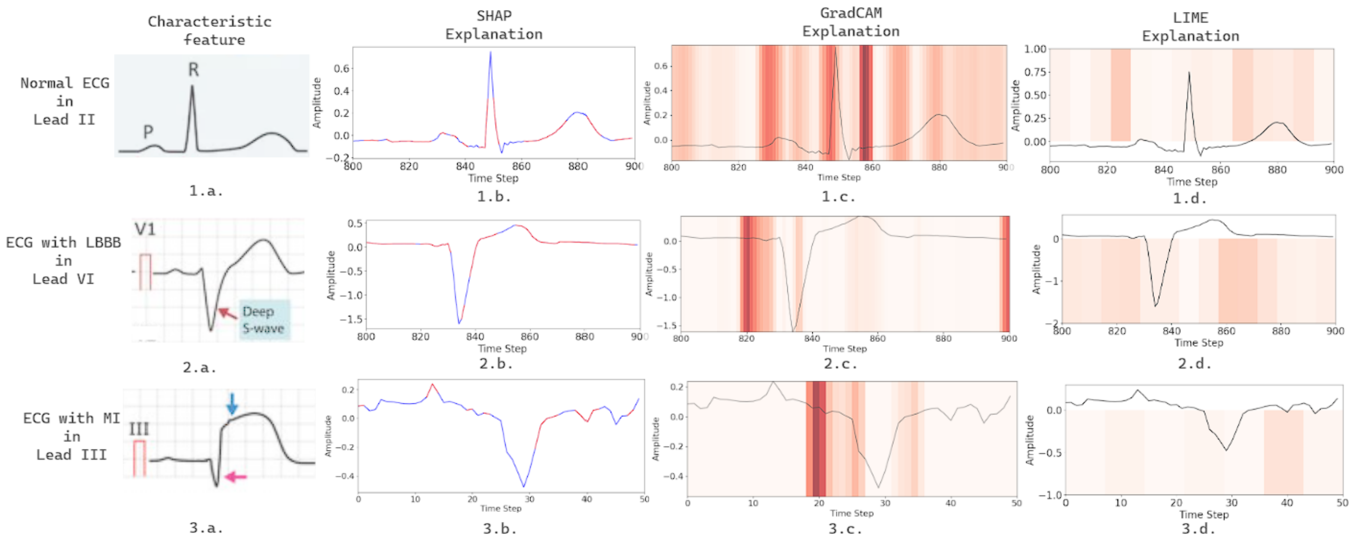

Figure 6: (Row 1) 1.a. Characteristic feature of a normal ECG in lead II [9]; 1.b. SHAP explanation of a normal ECG in Lead II; 1.c. Gradcam explanation of a normal ECG in Lead II; 1.d. LIME explanation of a normal ECG in Lead II. (Row 2) 2.a. Characteristic feature of LBBB in lead V1 [9]; 2.b. SHAP explanation of LBBB in Lead V1; 2.c. Gradcam explanation of LBBB in Lead V1; 2.d. LIME explanation of LBBB in Lead V1. (Row 3) 3.a. Characteristic feature of MI in lead III [9]; 3.b. SHAP explanation of MI in Lead III; 3.c. Gradcam explanation of MI in Lead III; 3.d. LIME explanation of MI in Lead III.

associated with ischemic heart disease when coupled with LBBB. LBBB falls under the broader superclass of Conduction Disturbance in the data set, emphasizing the model's ability to discern intricate patterns associated with various cardiac conditions.

In the context of explaining the model's decision for this instance, Gradcam and LIME exhibited inferior performance compared to SHAP, suggesting their limited suitability for interpreting our ECG classification model. The GradCAM heatmap predominantly highlights the PQ segment rather than the S segment of the conduction disturbance as seen in Figure 6.2.c.. Since GradCAM operates by computing gradients of the target class score concerning the feature maps of the last convolutional layer, it may not effectively capture the nuanced patterns associated with the S segment. The discrepancy observed in the LIME explanation likely stems from its method of generating local explanations through feature perturbation, which might struggle to capture the intricate relationships present in ECG data compared to SHAP's more comprehensive approach which considers the impact of all possible combinations of feature values, offering a more detailed understanding of the model's decision-making process.

Myocardial infarction (MI) commonly presents with pathological Q waves, which are wider and deeper than normal Q waves and ST-segment elevation in leads II, III, and aVF [9]. The report of the data set indicated sinus rhythm with borderline left axis deviation and the presence of such Q waves which are consistent with an old inferior myocardial infarction in leads II, III, and aVF. This classification was correctly identified by our model. Row 3 of Figure 6 depicts the explanation for Lead III.

While ST-segment elevation received significant attention from all three explanation models, it is noteworthy that pathological Q waves were not the primary focus of attention for myocardial infarction across these models. Despite being a characteristic feature of MI, these explanation models struggle to highlight their significance due to the complexity of ECG patterns. The presence of other ECG abnormalities alongside MI, coupled with the inherent challenges of interpreting medical data, could have further contributed to the limited focus on pathological Q waves in the explanation models.

5 CONCLUSION

In conclusion, this study reinforces the potential of deep learning models in accurately classifying various cardiac conditions from ECG data. The "ST-CNN-5 new" model outperforms existing models in terms of predictive accuracy. Its robustness and ability to handle complex ECG data make it a valuable asset for clinical applications.

Furthermore, the integration of interpretability techniques such as SHAP, GradCAM, and LIME provides valuable insights into the model's decision-making process, enhancing trust and understanding among healthcare professionals. While SHAP values offer valuable interpretability, we also acknowledge the limitations of other methods such as GradCAM and LIME. GradCAM may struggle with capturing fine-grained patterns, while LIME may not handle nonlinear relationships well. GradCAM predominantly focuses on the PQ segment rather than the S segment of conduction disturbances, while LIME struggles to emphasize the significance of pathological Q waves in myocardial infarction classification.

This research contributes to advancing the understanding of interpretable AI methods in ECG data classification, emphasizing the importance of both performance evaluation and interpretability analysis in model development.

Moreover, addressing challenges specific to ECG data, including temporal dependencies and noise, remains essential for further enhancing model performance and interpretability. To bridge these gaps, collaboration with domain experts such as cardiologists and clinicians is imperative to refine interpretability tools for practical use in real-world clinical settings. We can also experiment with explainability strategies that not only provide insights into model predictions but also enable meaningful interactions between clinicians and AI systems for future research.

ACKNOWLEDGMENTS

We gratefully acknowledge the Experimental Infrastructure for Exploration of Exascale Computing (eX3) server, provided by Simula Research Laboratory, for facilitating the model training processes conducted in this study, which is financially supported by the Research Council of Norway under contract 270053.

REFERENCES

[1] 2022. New Artificial Intelligence Tool Detects Often Overlooked Heart Diseases. https://www.cedars-sinai.org/newsroom/new-artificial-intelligence-tool-detects-often-overlooked-heart-diseases/. [Online; accessed on 19-February-2023].

[2] Amulya Agrawal, Aniket Chauhan, Manu Kumar Shetty, Mohit D Gupta, Anubha Gupta, et al. 2022. ECG-iCOVIDNet: Interpretable AI model to identify changes in the ECG signals of post-COVID subjects. *Computers in Biology and Medicine* 146 (2022), 105540.

[3] Keyvan Amini, Alireza Mirzaei, Mirtohid Hosseini, Hamed Zandian, Islam Azizpour, and Yagoob Haghi. 2022. Assessment of electrocardiogram interpretation competency among healthcare professionals and students of Ardabil University of Medical Sciences: a multidisciplinary study. *BMC Medical Education* 22, 1 (2022), 448.

[4] 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.

[5] Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Brad J Erickson, et al. 2019. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet* 394, 10201 (2019), 861–867.

[6] Abid Ali Awan. 2023. *Introduction to t-SNE.* https://www.datacamp.com/tutorial/introduction-t-sne

[7] Federico Cabitza, Davide Ciucci, and Raffaele Rasoini. 2019. A giant with feet of clay: On the validity of the data that feed machine learning in medicine. In *Organizing for the Digital World: IT for Individuals, Communities and Societies.* Springer, 121–136.

[8] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[9] ECG Waves. n.d.. *Clinical ECG Interpretation.* https://ecgwaves.com/course/the-ecg-book/

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press. http://www.deeplearningbook.org.

[11] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 2019. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 558–567.

[12] Steven A Hicks, Jonas L Isaksen, Vajira Thambawita, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Inga Strümke, Christina Ellervik, Morten Salling Olesen, et al. 2021. Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. *Scientific reports* 11, 1 (2021), 10949.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Jeremy Howard and Sylvain Gugger. 2020. Fastai: A layered API for deep learning. *Information* 11, 2 (2020), 108.

[15] Md Islam, Md Haque, Hasib Iqbal, Md Hasan, Mahmudul Hasan, Muhammad Nomani Kabir, et al. 2020. Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science* 1, 5 (2020), 1–14.

[16] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. 2020. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery* 34, 6 (2020), 1936–1962.

[17] Anthony H Kashou, Wei-Yin Ko, Zachi I Attia, Michal S Cohen, Paul A Friedman, and Peter A Noseworthy. 2020. A comprehensive artificial intelligence–enabled electrocardiogram interpretation program. *Cardiovascular Digital Health Journal* 1, 2 (2020), 62–70.

[18] Shaan Khurshid, Samuel Friedman, Christopher Reeder, Paolo Di Achille, Nathaniel Diamant, Pulkit Singh, Lia X Harrington, Xin Wang, Mostafa A Al-Alusi, Gopal Sarma, et al. 2022. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation* 145, 2 (2022), 122–133.

[19] Yogesh Kumar, Apeksha Koul, Ruchi Singla, and Muhammad Fazal Ijaz. 2023. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of ambient intelligence and humanized computing* 14, 7 (2023), 8459–8486.

[20] Hui Wen Loh, Chui Ping Ooi, Silvia Seoni, Prabal Datta Barua, Filippo Molinari, and U Rajendra Acharya. 2022. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine* (2022), 107161.

[21] Francisco Lopez-Jimenez, Zachi Attia, Adelaide M Arruda-Olson, Rickey Carter, Panithaya Chareonthaitawee, Hayan Jouni, Suraj Kapa, Amir Lerman, Christina Luong, Jose R Medina-Inojosa, et al. 2020. Artificial intelligence in cardiology: present and future. In *Mayo Clinic Proceedings*, Vol. 95. Elsevier, 1015–1039.

[22] Gianluca Malato. 2021. *How to explain neural networks using SHAP.* https://towardsdatascience.com/how-to-explain-neural-networks-using-shap-2e8a0d688730

[23] Sarang Narkhede. 2018. *Understanding AUC - ROC Curve.* https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[24] Ayush Pandey, Rakesh Chandra Joshi, and Malay Kishore Dutta. 2023. Automated Classification of Heart Disease using Deep Learning. In *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT).* IEEE, 358–362.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[26] 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, San Francisco, CA, USA, August 13-17, 2016.* 1135–1144.

[27] Konstantinos C Siontis, Peter A Noseworthy, Zachi I Attia, and Paul A Friedman. 2021. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology* 18, 7 (2021), 465–478.

[28] Sulaiman Somani, Adam J Russak, Felix Richter, Shan Zhao, Akhil Vaid, Fayzan Chaudhry, Jessica K De Freitas, Nidhi Naik, Riccardo Miotto, Girish N Nadkarni, et al. 2021. Deep learning and the electrocardiogram: review of the current state-of-the-art. *EP Europace* 23, 8 (2021), 1179–1191.

[29] Mayo Clinic Staff. 2022. Electrocardiogram (ECG or EKG). [Online; accessed on 19-February-2023].

[30] 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.

[31] Kush R. Varshney. 2022. *Trustworthy Machine Learning.* Independently Published. http://trustworthymachinelearning.com/.

[32] Neha Vishwakarma. 2023. *Visualizing Model Insights: A Guide to Grad-CAM in Deep Learning.* https://www.analyticsvidhya.com/blog/2023/12/grad-cam-in-deep-learning/

[33] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. 2020. PTB-XL, a large publicly available electrocardiography dataset. https://doi.org/10.13026/qgmg-0d46

[34] Zekai Wang, Stavros Stavrakis, and Bing Yao. 2023. Hierarchical deep learning with Generative Adversarial Network for automatic cardiac diagnosis from ECG signals. *Computers in Biology and Medicine* (2023), 106641.

[35] Zhiguang Wang, Weizhong Yan, and Tim Oates. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN).* IEEE, 1578–1585.

[36] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016).

[37] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency.* 295–305.