Information Sciences 665 (2024) 120383

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Application of spatial uncertainty predictor in CNN-BiLSTM model using coronary artery disease ECG signals

Silvia Seoni [a], Filippo Molinari [a,*], U. Rajendra Acharya [b], Oh Shu Lih [c], Prabal Datta Barua [d,e], Salvador García [f], Massimo Salvi [a]

[a] Biolab, PolitoBIOMedLab, Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy
[b] School of Mathematics, Physics and Computing, and Centre for Health Research, University of Southern Queensland, Springfield, Australia
[c] Cogninet Australia, Sydney, NSW 2010, Australia
[d] School of Business (Information System), University of Southern Queensland, Toowoomba, QLD 4350, Australia
[e] Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia
[f] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18071 Granada, Spain

ARTICLE INFO

Keywords:
Explainable AI
CAD
ECG
Deep learning
Uncertainty quantification
Signal processing

ABSTRACT

This study aims to address the need for reliable diagnosis of coronary artery disease (CAD) using artificial intelligence (AI) models. Despite the progress made in mitigating opacity with explainable AI (XAI) and uncertainty quantification (UQ), understanding the real-world predictive reliability of AI methods remains a challenge. In this study, we propose a novel indicator called the Spatial Uncertainty Estimator (SUE) to assess the prediction reliability of classification networks in practical Electrocardiography (ECG) scenarios. SUE quantifies the spatial overlap of critical Grad-CAM (Gradient-weighted Class Activation Mapping) features, offering a confidence score for predictions.

To validate SUE, we designed a deep learning network that integrates Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) mechanisms for precise ECG signal classification of CAD. This network achieved high accuracy, sensitivity, and specificity rates of 99.6%, 99.8%, and 98.2%, respectively. During test time, SUE accurately distinguishes between correctly classified and misclassified ECG segments, demonstrating the superiority of the proposed network over existing methods.

The study highlights the potential of combining XAI and UQ techniques to enhance ECG analysis. The evaluation of spatial overlap among discriminative features provides quantitative insights into the network's robustness, encompassing both current prediction accuracy and the repeatability of predictions.

1. Introduction

Coronary artery disease (CAD) is a major cardiovascular disorder affecting millions globally [1]. CAD primarily originates from atherosclerosis, where fibrous plaques develop in artery walls [2]. These plaques largely comprise fats and fibrous tissue [3]. CAD

* Corresponding author at: Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi, 24 – 10129 Turin, Italy.
E-mail address: Filippo.molinari@polito.it (F. Molinari).

https://doi.org/10.1016/j.ins.2024.120383
Received 30 October 2023; Received in revised form 31 January 2024; Accepted 28 February 2024
Available online 2 March 2024
0020-0255/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

manifestations range from asymptomatic to life-threatening myocardial infarction (MI). Severe cases involve plaque rupture triggering acute ischemic heart disease or MI [4]. Precise CAD diagnosis remains challenging due to varied presentations, necessitating advanced diagnostic tools. Electrocardiography (ECG) is the primary tool for its non-invasiveness and cost-effectiveness. Specific ECG signal changes are associated with ischemia, infarction, and arrhythmias [4]. ECG monitoring provides critical insights like ST-segment elevation indicating myocardial ischemia. However, nearly 70 % of CAD patients lack significant ECG changes [5], requiring exhaustive visual examination prone to fatigue-induced errors. Automated systems to address manual ECG limitations are essential [6].

Over the past decades, numerous signal processing algorithms and machine learning (ML) models have been developed and proposed for the automated detection of cardiac arrhythmias and abnormalities in ECG signals, such as those associated with CAD or MI [7,8,9,10]. Additionally, nonlinear signal processing techniques have also emerged as valuable tools for decoding ECG signals, including nonlinear dynamical systems, chaos theory, fractal analysis, and entropy-based methods [11,12]. Nevertheless, while demonstrating varying degrees of performance, conventional ML models reliant on handcrafted features have well-known inherent limitations stemming from their dependence on the quality, quantity, and representation of the training data.

In recent years however, there has been a notable increase in interest surrounding the novel utilization of deep learning (DL) models, such as convolutional and recurrent neural networks, for enhanced CAD detection directly from raw or minimally processed ECG signal data [13–17]. In contrast to conventional ML approaches, end-to-end DL models present a significant advancement in automating feature engineering and capturing intricate discriminative patterns within ECG morphological data. The rapid progress in technology, particularly in the processing capabilities of extensive ECG datasets, coupled with the increased accessibility of digitized ECG data, has led to the widespread adoption of DL techniques. This adoption is further driven by the profound clinical importance of improving CAD diagnosis. The application of DL methods in developing automated CAD detection systems holds promise for facilitating more precise, efficient, and scalable diagnostic procedures, thereby addressing certain challenges encountered by earlier methodologies.

Although AI models have achieved human-level performance, their application remains limited, primarily due to the lack of transparency in the models' prediction processes and the associated uncertainty. Ongoing research efforts are focused on increasing the interpretability of DL models and expanding their clinical utility. To address this challenge, researchers have turned to explainable artificial intelligence (XAI) techniques and uncertainty quantification (UQ) methods. The evolution in AI research follows a trajectory from optimizing model performance to addressing uncertainty [18–21]. Concurrently, there arose a need to make models more interpretable and explainable, leading to the development of XAI approaches.

XAI techniques refer to a set of methods used to enhance the transparency and interpretability of AI systems. These techniques aim to provide insights into how AI models make predictions or decisions, enabling humans to understand and trust the reasoning behind those outcomes. Using XAI methods like Gradient-weighted Class Activation Mapping (Grad-CAM) [22,23], it is possible to assess the confidence of the current prediction by generating qualitative heatmaps that highlight the most influential features contributing to the classification. Nevertheless, they fail to provide a comprehensive understanding of the complex relationships within these highlighted regions. To address this limitation, UQ techniques have been developed to estimate the model reliability by quantifying the uncertainty in the model's predictions [24]. These models offer a deeper understanding of the decision-making process, aiding clinicians in making informed decisions and improving communication with patients. However, UQ methods often require computationally intensive simulations or sampling techniques to quantify uncertainty. As a result, analyzing complex systems or large-scale models can be time-consuming and resource-intensive, limiting their practicality in real-time or high-dimensional scenarios.

With this knowledge, this work aims to develop a cost-effective computational framework leveraging AI that integrates elements of XAI with principles of UQ to compute the Spatial Uncertainty Estimator (SUE), a new indicator for assessing the repeatability of model predictions by quantifying the spatial congruence of the most influential features. The main contributions of this paper are:

1) We introduce a new reliability metric, called SUE, which quantifies the spatial congruence of influential features identified by the Grad-CAM during testing. This metric provides a measure of prediction accuracy. To the best of our knowledge, we are the first group to propose SUE for CAD ECG signals.
2) The SUE metric serves as a cost-effective and time-efficient quantitative link between the realms of XAI and UQ, offering a quantitative assessment of network reliability on a scale from 0 to 1.
3) We extensively validate the SUE by varying the noise overlaid on the ECG signal. The indicator consistently identifies the network that performs best under different noise levels.
4) The SUE exhibits a strong correlation with prediction accuracy, with higher values indicating correct classifications and lower values indicating misclassifications.
5) We validate the SUE in the task of classifying CAD in ECG signals. Additionally, we propose a classification network that integrates a CNN and Bidirectional Long Short-Term Memory (BiLSTM) mechanisms for accurate and reliable ECG signal classification, demonstrating superior performance compared to state-of-the-art methods and achieving excellent accuracy in CAD detection.

The rest of this paper is organized as follows: Section 2 presents an overview of the current approach of XAI/UQ for CAD; Section 3 provides an exhaustive description of the proposed indicator; Sections 4 and 5 report and discuss the experimental results.

2. Related works

In the research field of cardiac arrhythmia classification, XAI and UQ have played a crucial role, leading to notable contributions in

the literature. Table 1 reports the most influential and recent studies that have employed XAI or UQ techniques in ECG classification.

To address the need for more robust interpretability tools, Yoo et al. [25], Hughes et al. [26], and Maweu et al. [27] introduced distinct methods to interpret CNN models in arrhythmia detection. These methods include the Attention Branch Network (ABN) [25], Linear Interpretable Model-Agnostic Explanations (LIME) [30], and feature-extraction-based interpretability [27]. These interpretability approaches aim to make sense of predictions made by DL and CNN models in the context of ECG classification. Additionally, Rutger R. van de Leur [28] proposed Shapley Additive Explanations, focusing on enhancing interpretability in extreme gradient boosting decision tree (XGBoost) models. Incorporating the Grad-CAM technique, Varandas et al. [29] presented a DL model for cardiac arrhythmia classification, aiming to enhance model interpretability. Nevertheless, the heat maps generated by Grad-CAM methods frequently display variability, posing a challenge in extracting information about the model's reliability, especially during testing. Although these heatmaps facilitate a qualitative inspection of the signal features influencing predictions, they lack the provision of a quantitative measure for prediction reliability. Consequently, researchers have turned to UQ models, which quantify prediction uncertainty to facilitate reliability estimation. These models not only enhance decision-making for clinicians but also improve patient communication.

Park et al. [30] introduced a novel approach that combines a self-attention-based LSTM-FCN deep learning architecture with an ensemble model to enhance the accuracy of classifying six distinct arrhythmia types. In a contemporary study, Elul et al. [31] presented the Monte Carlo Dropout (MCD) method to quantify uncertainty in cardiac arrhythmia classification, providing insights into prediction reliability. Barandas et al. [32], Vranken et al. [33], and Asseri et al. [33]conducted a comparative study for UQ techniques (MCD and Ensemble model) in cardiac arrhythmias detection. Notably, Barandas et al. [32] highlighted the robustness of ensemble methods in managing UQ and calibration amid dataset shifts. Their study revealed that ensemble-based methods outperformed single-network or stochastic methods in performance. The incorporation of uncertainty estimates into the classification process significantly improved the model's ability to adapt to shifts in data distribution. The study also emphasized the importance of external validation in multi-label ECG classification, an aspect that is often overlooked.

Jahmunah et al. [14] presented a Dirichlet DenseNet model for the MI classification in ECG signals. They employed predictive entropy as a reliable measure of uncertainty, enabling the detection of misclassifications between normal and MI ECG signals. Belen et al. [40] implemented test-time augmentation for UQ, adopting a unique approach. Lastly, Zangh et al. [34] introduced a Bayesian network with MCD for arrhythmia detection, emphasizing the importance of total uncertainty computation through data and model uncertainty decomposition. Their exploration of different uncertainty thresholds served to enhance classification performance by identifying and rejecting high-uncertainty samples.

It is important to acknowledge that XAI and UQ techniques in ECG analysis have limitations. XAI techniques often struggle to generalize qualitative information extracted, hindering comparisons across different conditions. The accuracy of UQ depends on factors such as model choice, data quality, and UQ methods. Despite advancements in XAI and UQ methods to address deep learning

Table 1
List of works done on XAI and UQ using ECG signals.

Authors, year	Dataset	XAI/ UQ	XAI/UQ technique	Aim
Barandas et al. [32], 2023	CRBBB in G12EC and PTB-XL dataset	UQ	MCD and Laplace approximation, ensemble methods	Cardiac arrhythmias detection
Jahmunah et al.[14], 2023	PTB-XL	UQ	Predictive entropy for model uncertainty	MI classification
Park et al. [30], 2023	MIT-BIH and INCART	UQ	Deep ensemble approach	Six arrhythmia classification
Rutger R. van de Leur [28], 2022	Private dataset	XAI	FactorECG (interpretable statistic model)	Detection of variation in ECG signals
Varandas et al.[29], 2022	MIT-BIH arrhythmia database	XAI	Grad-CAM	Arrhythmia classification
Zhang et al, [34], 2022	CPSC 2018	UQ	Bayesian neural network with MCD	Arrhythmia classification
Asseri et al. [33], 2021	MIT-BIH and INCART	UQ	MCD and deep ensemble method	Cardiac arrhythmias classification
Elul et al. [31], 2021	MIT-BIH	UQ	MCD	Heterogeneous mix of known and unknown arrhythmia detection
Hughes et al. [26], 2021	365 009 patients	XAI	LIME	38 arrhythmia ECG diagnoses
Maweu et al. [27], 2021	MIT-BIH	XAI	Explainability based on the features extraction	abnormal ECG classification
Vranken et al. [33], 2021	UMCU-Triage UMCU-Diagnose CPSC2018	UQ	MCD, variational inference, and ensemble	ECG Classification
Yong-Yeon Jo et al. [25], 2021	PTB-XL ECG; Georgia ECG challenge; CPSC ECG	XAI	Attribution maps	Arrhythmia detection
Yoo et al. [25], 2021	CPSC 2018 dataset	XAI	ABN	Arrhythmia classification
Belen et al. [35], 2020	MIT-BIH Atrial Fibrillation database	UQ	UQ estimation using test-time augmentation	Atrial fibrillation classification

*MCD: Monte Carlo dropout, LIME: Linear Interpretable Model-Agnostic Explanations, ABN: Attention Branch Network, MIT-BIH: Massachusetts Institute of Technology - Beth Israel Hospital dataset, CPSC: China Physiological Signal Challenge, UMCU-Triage and UMCU-Diagnose: two datasets acquired from University Medical Center Utrecht, PTB-XL: A Large Benchmark Dataset for PhysioNet/Computing in Cardiology Challenge 2020; INCART: 12-Lead Arrhythmia Database.

model opacity, there is still a noticeable gap in understanding their predictive reliability during real-world testing.

3. Materials and methods

In our study, we introduced a novel AI framework for quantifying the reliability of a network's prediction of CAD from ECG signals. CAD detection was performed by a CNN model integrated with BiLSTM layers. During testing, the ECG signals are corrupted by different types of noise at increasing power levels. For each pair of original and corrupted signals, the spatial overlap of features extracted by GRAD-CAM is computed, referred to as the SUE. The SUE provides quantitative information about where the network extracts the most relevant features to classify a corrupted signal compared to its baseline (original signal). Fig. 1 illustrates the proposed framework for assessing the reliability of a network's CAD classification of ECG signals.

3.1. Dataset

In this study, we sourced healthy ECG signals from the Physionet databases, specifically the *Fantasia* dataset [36]. These signals had a recording duration of 120 min in Lead II, with a sampling frequency of 250 Hz. For the CAD ECG signals, we used 75 signals extracted from 32 Holter monitor recordings from the St.-Petersburg Institute of Cardiology Technics 12-lead arrhythmia dataset [36]. The CAD signals were recorded at a sampling frequency of 257 Hz, with each signal lasting 20 min. To confirm the presence of CAD, comprehensive enzymatic assays, coronary angiography, electrophysiological studies, and vigilant blood pressure monitoring were conducted.

In total, our study included healthy ECG signals from 40 individuals without cardiac irregularities (50 % females) and ECG signals from 7 individuals diagnosed with CAD (1 male and 6 females). To ensure compatibility with the CAD signals, we downsampled the healthy ECG signals to 257 Hz. Additionally, we applied a discrete wavelet transform (DWT) using the Daubechies 6 (6 db) wavelet to mitigate noise artifacts and rectify baseline aberrations [37]. Subsequently, the pre-processed signals were segmented into 2-second epochs, resulting in a total of 514 samples for each epoch. The final dataset consisted of a total of 95,300 ECG segments, categorized into 80,000 healthy ECG segments and 15,300 pathological CAD ECG segments.

For the model training process, the dataset was divided at the patient level in training, validation and test set, as follows: 60,039 segments (n = 29) were allocated for training, 25,731 (n = 13) for validation, and 9530 (n = 5) for testing purposes. Table 2 presents the dimensions of the dataset, including both classes, across the training, testing, and validation sets.

3.2. CNN + BiLSTM network

The proposed model architecture is designed with a sequence of three 1D convolutional (1D Conv) layers, followed by two BiLSTM layers [38], and three subsequent Dense layers. BiLSTM models are an extension of Recurrent Neural Networks (RNNs) and offer an effective solution to the vanishing gradient problem [39]. Deep-bidirectional LSTMs [38] further enhance the capabilities of LSTM models by applying two LSTMs to the input data. The utilization of BiLSTM enables the model to capture long-term dependencies and enhances overall accuracy [40].

The first two Dense layers employ the Rectified Linear Unit (ReLU) activation function, while the final layer utilizes the Softmax activation function to generate predicted probabilities for the two distinct classes. The input of the model was the ECG segments, while

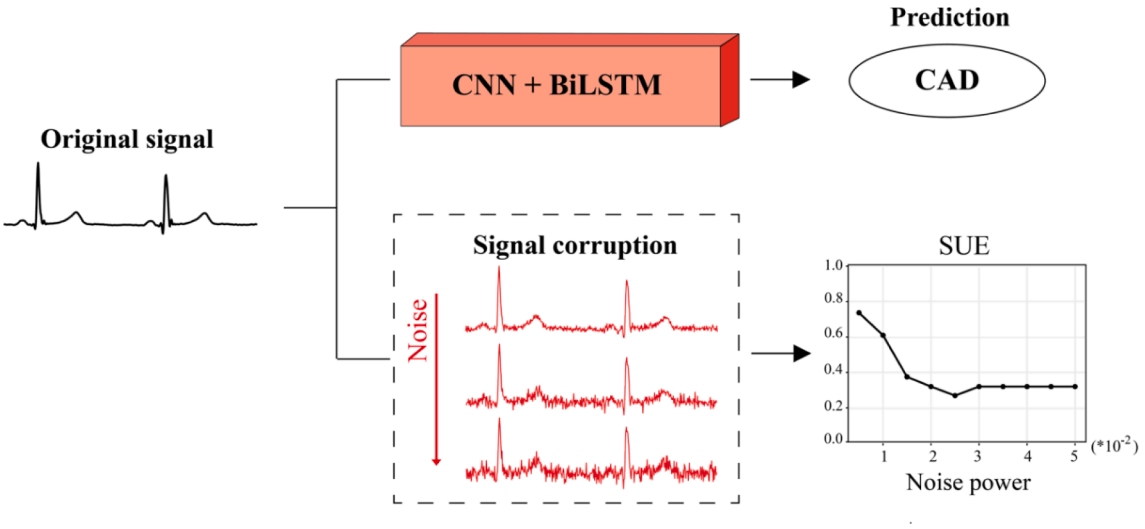


Fig. 1. The proposed AI framework assesses the reliability of a network's prediction during test time. During the test time, the ECG signals were corrupted using different types of noise, and a new indicator SUE is computed to assess the robustness of the current prediction.

Table 2
Dimension of the training, validation, and test set.

	Healthy	CAD	Total
Training set (n = 29)	50,407	9632	60,039
Validation set (n = 13)	21,544	4187	25,731
Test set (n = 5)	8049	1481	9530

the output was CAD predictions. The proposed architecture is illustrated in Fig. 2, and a comprehensive overview of its design is provided in Table 3. During the training process, we employed the Adam optimization method and used the Binary Cross-Entropy loss function. Building upon insights from previous studies [41], the proposed model was trained for 20 epochs with a batch size of 10.

3.3. Dataset corruption

In real scenarios, ECG signals are susceptible to contamination from various noise sources, including motion artifacts, muscle noise, and baseline wander, each exhibiting unique characteristics and properties [42]. With this knowledge, we incorporate three distinct types of noise to corrupt the original ECG dataset during the testing phase. To quantify the level of signal corruption, we measure the power of the noise, which directly affects the signal-to-noise ratio (SNR) [43]. We employ four different noise sources: synthetic Gaussian white noise, synthetic power line noise, and two real noise records from the PhysioNet noise stress database [44].

Firstly, we synthesize corrupted ECG signals by incorporating Gaussian white noise samples with varying power levels ranging from 0.001 to 0.005. Next, we synthesized signals with a real-world type of ECG noise: power line interference. The ECG data was corrupted by the synthesized noise, with the noise power spanning a range from 0.001 to 0.01. Additionally, we use two distinct real noise types extracted from ECG signals (Leads I and II) acquired from subjects' limbs [44]: *em* and *ma*. The *em* records exhibit noise accompanied by substantial baseline wander, while the *ma* records primarily contain muscle noise. The corruption process involves introducing noise with variable power, spanning a range from 0.0001 to 0.0005. Fig. 3 shows an example of healthy and CAD signals along with their corresponding corruptive signals from the test set. In total, we constructed 20 different corrupted datasets:

1) 5 datasets with Gaussian noise-corrupted ECG data (ECG_{GAUSS}), with noise power ranging from 0.001 to 0.005.
2) 5 datasets with power line noise-corrupted ECG data (ECG_{LINE}), with noise power ranging from 0.001 to 0.01.
3) 5 datasets with '*em*' noise-corrupted ECG data (ECG_{EM}), with noise power ranging from 0.0001 to 0.0005.
4) 5 datasets with '*ma*' noise-corrupted ECG data (ECG_{MA}), with noise power ranging from 0.0001 to 0.0005.

The original ECG data, without any noise corruption, will be referred to as ECG_{ORIG}.

3.4. Spatial uncertainty estimation (SUE)

The introduction of a novel metric called SUE can facilitate the evaluation of AI model reliability in real-world testing scenarios. SUE is specifically designed to assess the repeatability of model predictions by quantifying the degree of overlap among the most influential features identified through Grad-CAM-generated heatmaps. Fig. 4a presents the proposed framework for evaluating

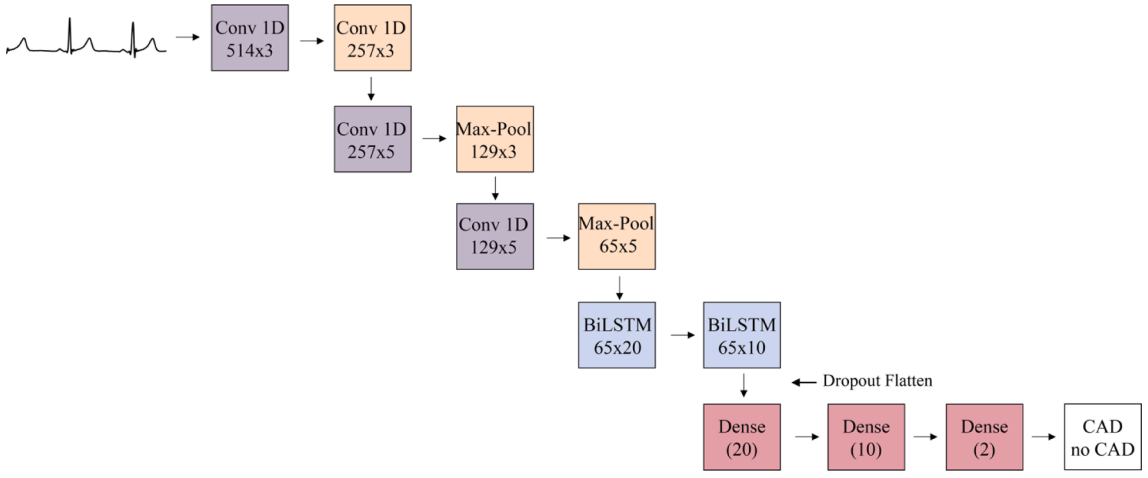


Fig. 2. Overview of the proposed architecture: CNN + BiLSTM. The architecture comprises three 1D convolutional layers (Conv 1D), followed by two Bidirectional Long Short-Term Memory (BiLSTM) layers, and three subsequent Dense layers. The input consists of ECG signals (512 samples), and the output is the prediction of 'CAD' or 'no CAD'.

Table 3
The proposed architecture: layers, output and kernel dimension, and stride.

Layer	Output dimension	Kernel dimension	Stride
Conv 1D	514x3	27	1
Max Pooling	257x3	2	2
Conv 1D	257x5	15	1
Max Pooling	129x5	2	2
Conv 1D	129x5	4	1
Max Pooling	65x5	2	2
BiLSTM	65x20	–	–
BiLSTM	65x10	–	–
Fully connected	20	–	–
Fully connected	10	–	–
Fully connected	2	–	–

prediction reliability during the test phase, which involves three distinct stages: (1) heatmap estimation, (2) heatmap discretization, and (3) SUE computation.

Initially, the ECG_{ORIG} dataset is used to test the CNN-BiLSTM model, and the Grad-CAM technique is employed to estimate the heat maps. The Grad-CAM is applied to the final BiLSTM layer. The heatmaps exhibit considerable variability, posing challenges in assessing reliability predominantly through qualitative means. To enable more effective comparisons, in this study, the heatmaps are discretized into a binary representation using percentile values as criteria. Specifically, the heatmaps are categorized into four distinct classes based on their values at the 25th, 50th, and 75 ^{h} percentiles:

- Very low influence features: values below the 25th percentile.
- Low influence features: values within the range [25th, 50th percentile].
- Moderate influence features: values within the range [50th, 75th percentile].
- High relevant features: values above the 75th percentile.

To focus exclusively on the most relevant portions of the heatmap, a further discretization step is implemented to create a binary classification task:

1) Class *High* (representing high-relevant features): values above the 75th percentile.
2) Class *Low* (representing low-influence features): values below the 75th percentile.

Subsequently, the same procedure is applied to the corrupted ECG datasets (ECG_{GAUSS} ECG_{EM} ECG_{MA}). Considering an epoch of ECG_{ORIG} and ECG_{GAUSS}, the SUE is computed as the intersection over the union of the original and corrupted heatmap for Class *High*, as described in Equation (1):

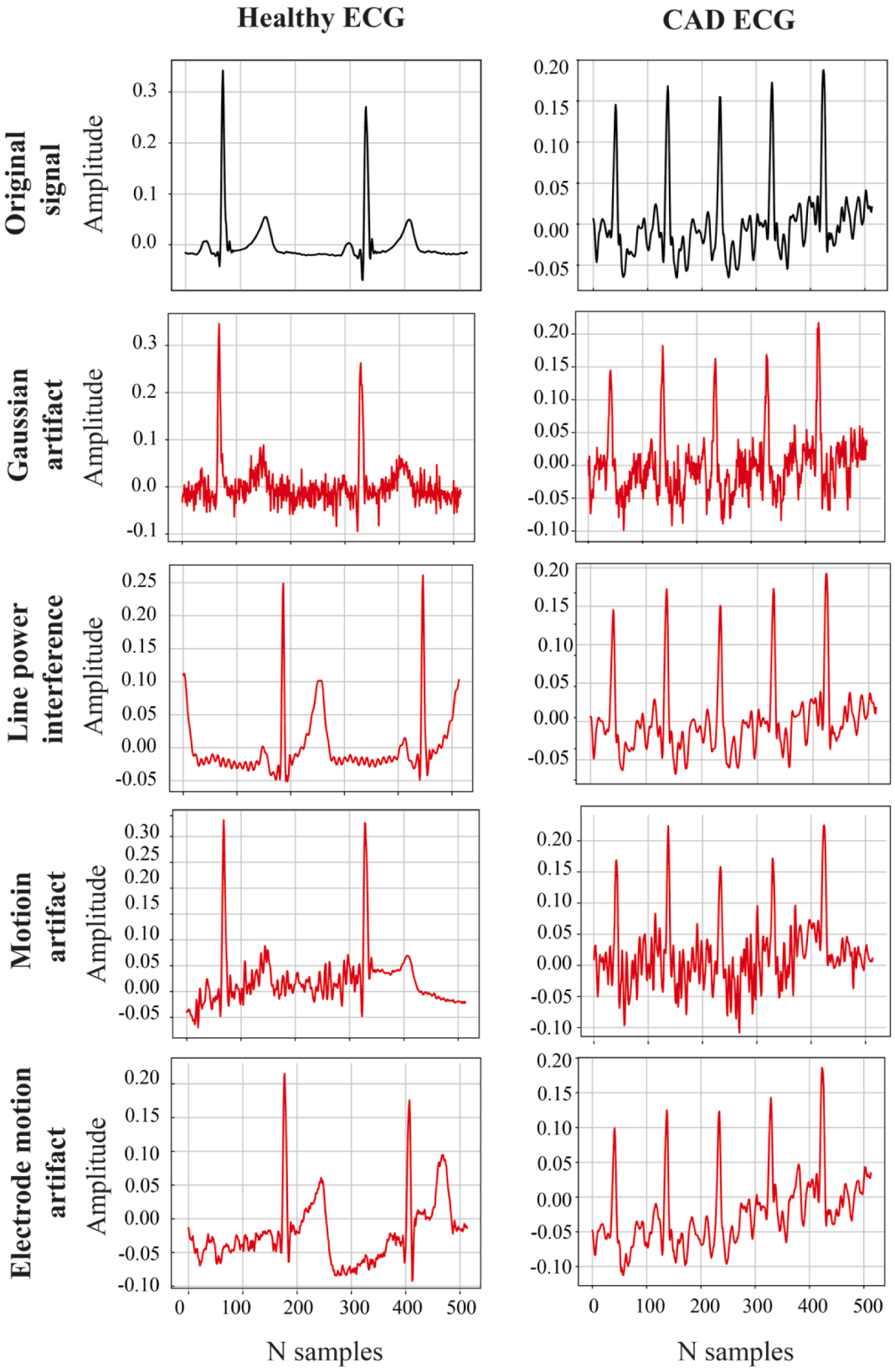
$$SUE = \frac{heatmap_{original}^{[High]} \cap heatmap_{corrupted}^{[High]}}{heatmap_{original}^{[High]} \cup heatmap_{corrupted}^{[High]}} \tag{1}$$

A SUE value close to 1 demonstrates a high level of reliability in the prediction, indicating a nearly perfect spatial overlap between the relevant features in the original signal and their corresponding features in the corrupted signal. Conversely, a lack of prediction reliability is indicated by a SUE tending towards 0, suggesting that the spatial concurrence of the relevant features cannot be discernible. In summary, a SUE value of 1 implies a close resemblance between the relevant features identified in the original heatmap and those in the corrupted heatmap, while a dissimilarity in the high-relevant regions of the heatmaps in the two scenarios leads to a SUE value approaching zero.

Furthermore, extensive validation of the SUE is conducted by introducing varying levels of noise to the ECG signal. This allows for the assessment of prediction repeatability under different conditions, such as an increased noise power level in the corrupted signal. Fig. 4b illustrates the computation of SUE between the original epoch and corrupted epochs with different noise power levels, and the quantitative trend of prediction reliability.

3.5. Performance metrics

During testing, we first evaluate the performance of our proposed model (CNN-BiLSTM) compared to widely used popular networks (CNN and DenseNet) in terms of accuracy, sensitivity, and specificity. The first model (CNN) was specifically developed for CAD signal classification and exhibited outstanding performance using the same dataset employed in our study [41]. The architecture of the CNN model consists of four Conv1D layers, interspersed with a max-pooling layer and three final Dense layers. The second model (DenseNet) was designed for the classification of myocardial infarction (MI) in ECG signals and integrated with the UQ [14]. The DenseNet model comprises 5 layers with 1D, followed by 1D average pooling, 9 layers with 1D followed by 1D average pooling, global average pooling, and a fully connected layer.



(caption on next page)

Fig. 3. Example of healthy (left) and CAD-affected (right) ECG signals, along with their respective corrupted signals from the test set. The top row displays the original signals. Subsequent rows show signals corrupted by Gaussian noise, power line interference, motion artifacts, and electrode motion artifacts, respectively.

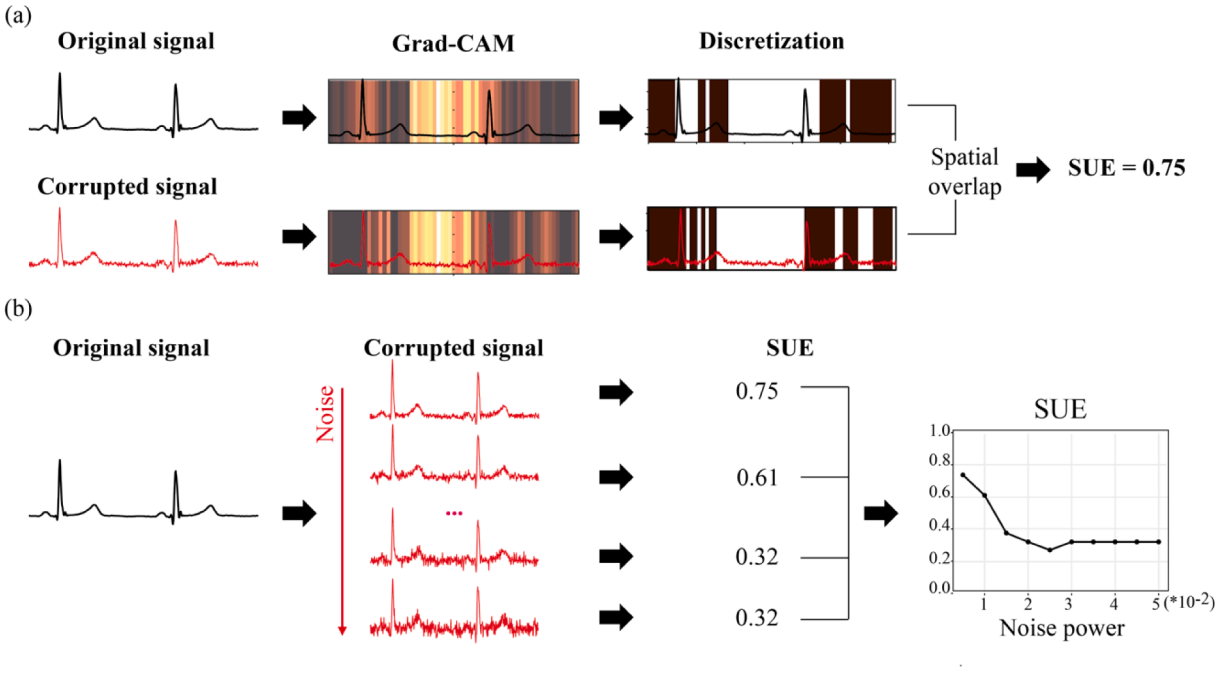


Fig. 4. The proposed AI framework for reliability assessment using the SUE parameter. (a) First, heatmaps are estimated from the original and corrupted signals. Then, after heatmap discretization, the most relevant features are retained. Finally, the spatial overlap is computed. (b) An example of extended SUE validation, achieved by varying the noise overlaid on the ECG signal.

To estimate the SUE, for both CNN and DenseNet models, Grad-CAM is applied to the final convolutional layer. Following the validation of these networks, we proceed to conduct an extensive evaluation of the SUE metrics. This evaluation serves a dual purpose: firstly, to identify the best-performing network during testing, and secondly, to assess the reliability of predictions. SUE not only identifies the most accurate model but also offers valuable uncertainty measurements, making it a useful tool for assessing the reliability of predictions.

We first compare prediction repeatability with SUE values derived from the three models under various levels of ECG signal corruption. This facilitates a direct comparison of prediction and feature repeatability, optimizing the quantification of prediction robustness.

Furthermore, we investigate the relationship between SUE values and prediction accuracy by separately estimating SUE for correctly and incorrectly classified instances. This evaluates whether SUE correlates with model accuracy. This comparative evaluation is performed for each noise power level applied to corrupt ECG signals.

Finally, we compared the SUE to other established uncertainty estimation methods in the literature: Deep Ensemble and Monte Carlo Dropout (MCD) [45]. To evaluate Deep Ensemble, we developed a model using the three trained networks - CNN, DenseNet, and CNN-BiLSTM. We then measured the uncertainty of predictions on the test set using Deep Ensemble. This allowed us to compare SUE to

Table 4
Model performance on the train, validation, and test set: CNN, DenseNet and the proposed model (CNN-BiLSTM).

Network	Subset	Accuracy	Sensitivity	Specificity
CNN [41]	train	0.991	0.994	0.977
	validation	0.988	0.993	0.963
	test	0.988	0.993	0.963
DenseNet [14]	train	0.989	0.992	0.967
	validation	0.985	0.991	0.957
	test	0.986	0.991	0.959
CNN-BiLSTM (proposed)	train	**0.999**	**0.999**	**0.992**
	validation	**0.997**	**0.998**	**0.988**
	test	**0.996**	**0.998**	**0.982**

other common uncertainty quantification techniques.

4. Results

4.1. Classification performance

The validation parameters for the proposed model, as well as the CNN and DenseNet models, are presented in Table 4. Throughout both the validation and testing phases, the CNN-BiLSTM model consistently demonstrates superior classification performance. It demonstrated superior accuracy, sensitivity, and specificity in correctly identifying and classifying cardiac arrhythmias. In the test set, its accuracy reaches 0.996, while the DenseNet and CNN achieve 0.986 and 0.988, respectively. However, it is important to note that the CNN and DenseNet models also displayed excellent performance, indicating their effectiveness in ECG signal classification. Fig. 5 shows the training curves for all the tested models, displaying loss and accuracy.

To provide a visual representation of the model's performance, Fig. 6 presents the confusion matrices (CMs) obtained using the test set. Considering the CMs, the number of misclassifications is consistently higher in both the CNN and DenseNet networks compared to our proposed CNN-BiLSTM model. Specifically, in the CAD class, the CNN-BiLSTM model demonstrates a notable decrease in the misclassification of non-CAD segments as CAD (False Positives), with only 13 epochs compared to 57 and 68 epochs for the CNN and DenseNet models, respectively. Additionally, our proposed model exhibits a significant reduction in misclassifying CAD segments as non-CAD (False Negatives), with 26 segments compared to 55 and 61 segments for the CNN and DenseNet models, respectively. This outcome holds clinical significance as it contributes to minimizing the occurrence of False Negatives, which is crucial for accurate diagnosis and effective treatment.

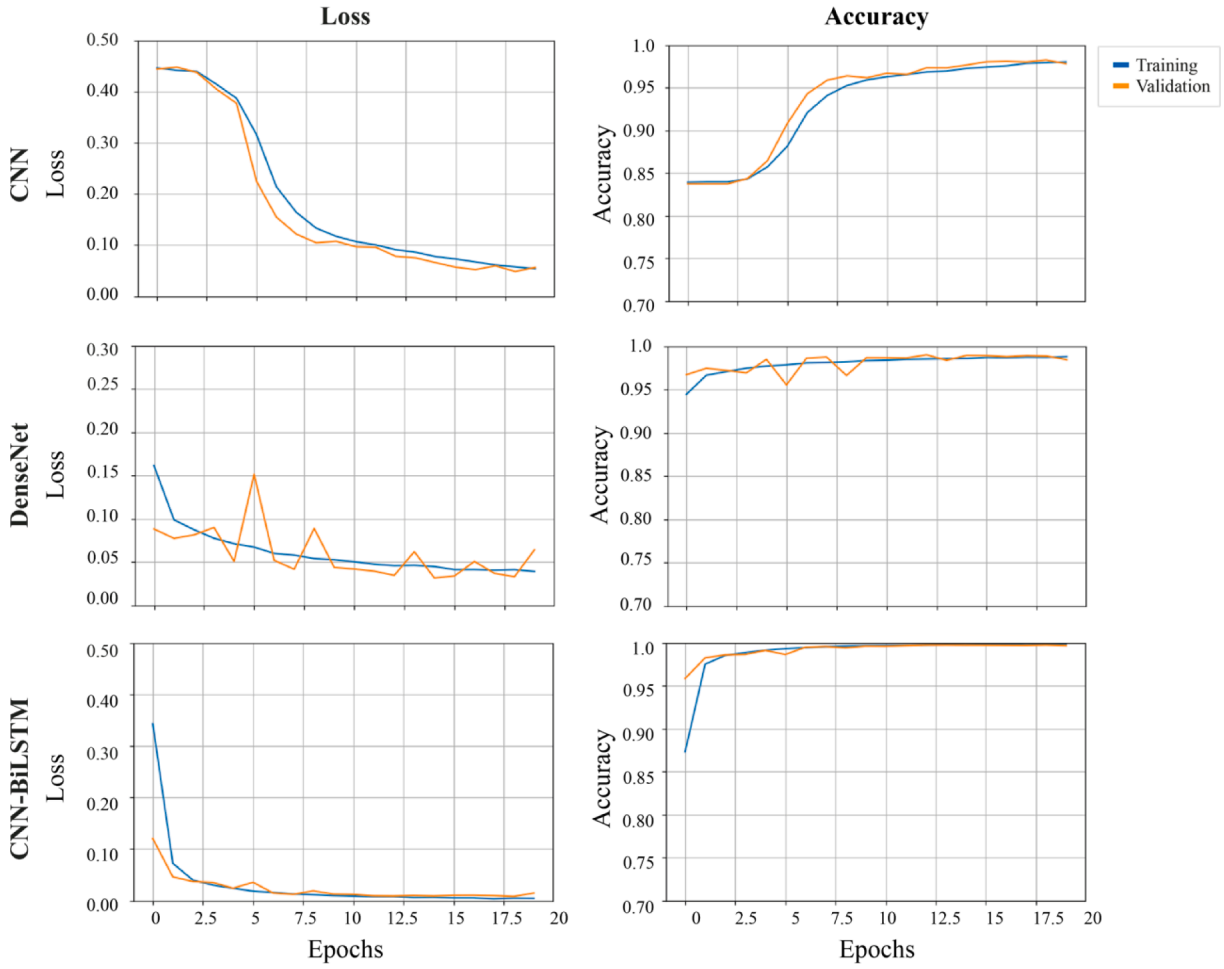


Fig. 5. Training curves for the three tested models, displaying loss in the first column and accuracy in the second column. Loss and accuracy are depicted for the CNN model on the first line, for DenseNet on the second line, and for CNN-BiLSTM on the last line.

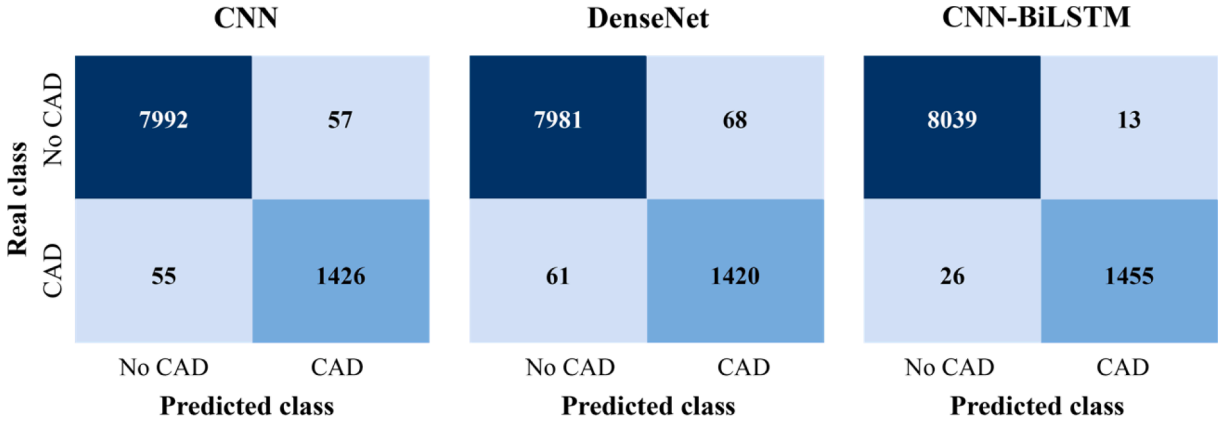


Fig. 6. The confusion matrices of the CNN, DenseNet, and the proposed model CNN-BiLSTM, estimated using the test set.

4.2. SUE validation

Typically, prediction repeatability is the conventional metric used to evaluate model reliability during testing. However, this metric often fails to provide meaningful insights into prediction reliability, especially when predictions exhibit systematic errors or misclassifications. SUE metric can be used to quantify the reliability of model predictions by assessing the spatial overlap between the most influential features through heatmaps generated using the Grad-CAM technique. In Fig. 7, we compare SUE with the prediction repeatability for the CNN-BiLSTM, CNN, and DenseNet models across the four different noise conditions: Gaussian white noise (first line), power line interference (second line), motion artifact noise (third line), and electrode motion artifact (fourth line).

In the first analysis, we evaluated the reliability of predictions by introducing Gaussian noise at different power levels ranging from 0.001 to 0.005. Overall, both the CNN and CNN-BiLSTM models performed the best. When considering the repeatability of predictions alone, the two models showed similar reliability. However, as we increased the power of the noise, a noticeable contrast emerged when examining the SUE values. The CNN-BiLSTM model proved to be more reliable, reinforcing its superior performance as also highlighted in Table 4. Although DenseNet exhibited commendable performance under normal conditions, it demonstrated reduced robustness when dealing with signals containing overlapping noise.

In the second analysis, we assess the robustness of the three models using the corrupted ECG_{LINE}, synthesized using the power line interference varying the power noise in a range from 0.001 to 0.01. The proposed CNN-BiLSTM and the CNN model demonstrate comparable repeatability of predictions for almost all levels of noise power. However, when evaluating the SUE value, the CNN proves to be more robust than the CNN-BiLSTM for high levels of noise power (greater than 0.005).

In the third analysis, we evaluated the reliability of the three models using the ECG_{MA} dataset while varying the muscle noise overlaid on the ECG signal from 0.001 to 0.005. When considering only the repeatability of predictions, there were no evident differences between the CNN and CNN-BiLSTM models. However, when we considered SUE, a more pronounced distinction between these models became apparent. Furthermore, even in the presence of motion artifact noise, DenseNet showed the lowest level of robustness among the three models.

In the final analysis, we assessed the reliability of predictions using synthesized signals containing electrode motion artifacts. When considering the repeatability of predictions, both the CNN and CNN-BiLSTM models showed comparable performance, which was higher than that of the DenseNet model, even at low noise levels. However, when examining the SUE metric, a clear distinction between the two networks emerged from the beginning, with the CNN-BiLSTM model proving to be the most reliable, even in comparison to the CNN model.

4.3. SUE vs. Classification accuracy

In this section, we explore the relationship between SUE values and prediction accuracy. Fig. 8 displays the SUE computed for the correctly classified and misclassified ECG segments of the test set. The first column of Fig. 8 illustrates the SUE values for correctly and incorrectly classified cases at the initial low noise power level. From the same figure, it can be observed that the SUE is consistently higher for the correctly classified compared to the misclassified, especially in real-world noises.

To further investigate, we extend the analysis by calculating SUE values across increasing noise levels. The second column of Fig. 8 displays the mean SUE values and standard error estimates for different noise intensities for the correctly classified (blue lines) and misclassified (orange lines). This enables an assessment of the trend in SUE for correctly classified and misclassified cases as the noise power varies. Consistently, SUE values remain higher for correctly classified ECG segments compared to misclassified segments as noise levels increase.

These findings establish a robust positive correlation between SUE values and classification accuracy during the test phase. Higher SUE values correspond to more accurate classification, while lower SUE values indicate a higher number of misclassifications. Importantly, this relationship holds true even as noise levels intensify.

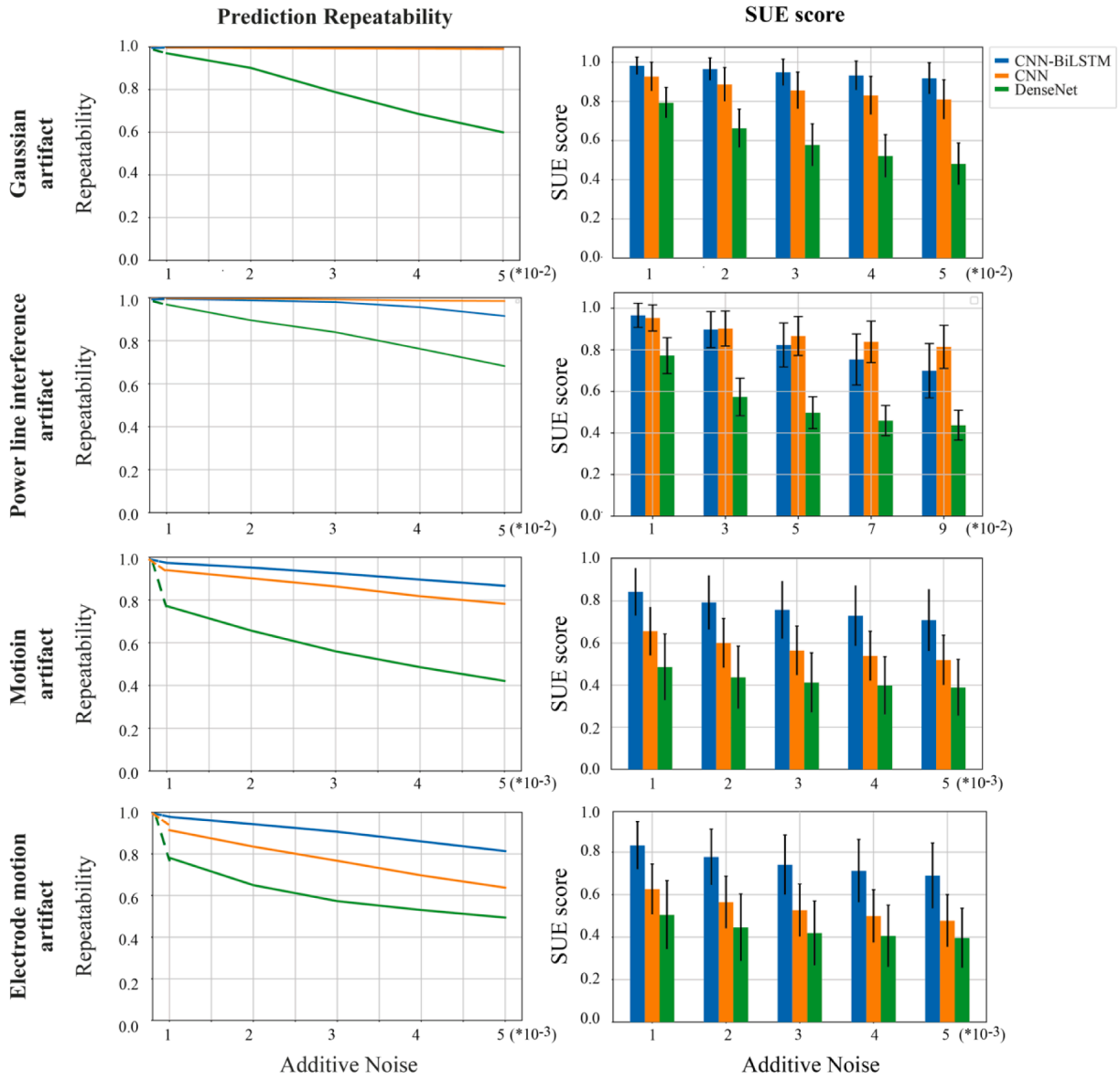


Fig. 7. Comparison between prediction repeatability and SUE scores during test time. The first column illustrates the trend in prediction repeatability as noise power levels are varied to corrupt ECG signals. The blue line corresponds to the proposed model (CNN-BiLSTM), the orange line represents the CNN, and the green line represents the DenseNet. In the second column, the mean values and standard deviations of SUE are presented for different noise power levels applied to distort the ECG signals. The blue graph represents SUE values estimated using our method (CNN-BiLSTM), the orange graph represents values estimated with the CNN, and the green graph represents values estimated with the DenseNet.

4.4. SUE vs. Other UQ methods

The SUE metric was compared to uncertainty estimation methods found in the literature, such as Deep Ensemble and MCD. Uncertainty values were estimated using the test set, which was corrupted by Gaussian noise, line interference, movement artifact, and electrode motion artifact. The noise power values were set to 0.001 for synthetic noises and 0.0001 for real noises. These uncertainty values were then compared to the SUE value estimated under the same conditions. To facilitate comparison, the SUE is presented as 1-SUE, which provides information about the prediction uncertainty.

Table 5 presents the mean value and standard deviation of uncertainty estimated using Deep Ensemble, CNN-BiLSTM with MCD, and CNN-BiLSTM with SUE. Uncertainty values are estimated for both correctly classified (CC) and misclassified (MC) instances. As shown in the table, the methods capable of providing distinct values for CC and MC instances are the Deep Ensemble and SUE. A notable difference of around 30 % can be observed between the uncertainty values produced by the Deep Ensemble for correctly classified test samples and misclassified samples. Similarly, the SUE demonstrates significant differentiation between CC and MC

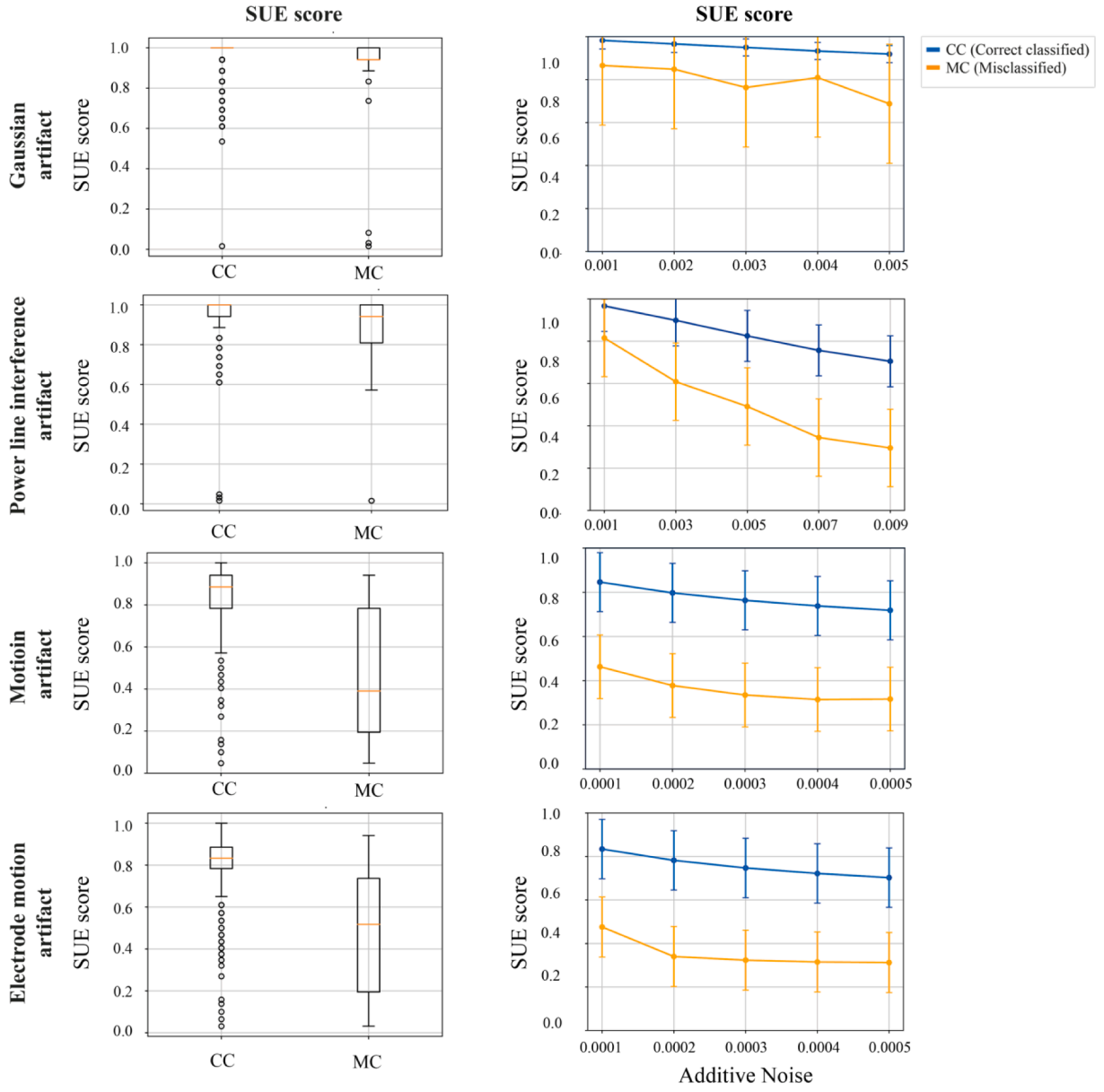


Fig. 8. SUE scores estimated for correctly classified and misclassified ECG segments. The right column displays SUE values under lower noise power conditions (0.001 for Gaussian noise and power line interference, while 0.0001 for real artifact noise), for the correctly classified (CC) and misclassified (MC). The left column presents mean values and standard errors of SUE scores estimated with various noise powers. The blue line corresponds to SUE values for correctly classified segments, and the orange line corresponds to those for misclassified segments.

Table 5
Comparison between the Uncertainty estimated in the Deep Ensemble model, CNN-BiLSTM + Monte Carlo Dropout and CNN-BiLSTM + SUE for all the corrupted ECG datasets.

Techniques		Gaussian artifacts	Power line interference	Motion artifacts	Electrode artifacts
Deep Ensemble	CC	0.011 ± 0.058	0.011 ± 0.059	0.029 ± 0.095	0.029 ± 0.096
	MC	0.361 ± 0.132	0.355 ± 0.139	0.350 ± 0.149	0.339 ± 0.148
CNN-BiLSTM + MCD	CC	0.001 ± 0.003	0.001 ± 0.002	0.001 ± 0.005	0.001 ± 0.004
	MC	0.027 ± 0.041	0.032 ± 0.041	0.032 ± 0.040	0.020 ± 0.033
CNN-BiLSTM + SUE	CC	0.033 ± 0.054	0.017 ± 0.039	0.166 ± 0.107	0.153 ± 0.106
	MC	0.184 ± 0.290	0.134 ± 0.277	0.524 ± 0.294	0.537 ± 0.288

*CC represents the correctly classified and the MC represents the misclassified segments.

instances. Specifically, for the two types of real noise (motion artifact and electrode artifact), the SUE exhibits a difference of approximately 40 %. However, MCD does not yield significant differences between correctly classified and misclassified samples in the test set.

5. Discussion

CAD is a prevalent cardiovascular disorder affecting millions of individuals. Early diagnosis is vital for improving treatment outcomes. Automated diagnostic systems are crucial in achieving early detection and overcoming the limitations of manual ECG assessments. Despite achieving human-level performance, the practical use of AI models is constrained by transparency issues and inherent uncertainty. Ongoing research endeavors focus on enhancing the interpretability of DL models and expanding their clinical applicability, prompting the integration of XAI techniques and UQ methods. Despite progress in mitigating opacity with XAI and UQ, understanding their real-world predictive reliability remains a challenge. Indeed, the XAI provides insights, but it struggles to generalize information and make comparisons between different conditions. On the other hand, UQ accuracy depends on factors such as model selection, data quality, and specific methodologies employed. Therefore, there is a need for a robust indicator to quantify prediction resilience in practical ECG scenarios. In this study, we presented an AI-based framework that computes the SUE, a novel indicator for assessing the prediction reliability of classification networks during testing.

The SUE indicator is developed to overcome the limitations of traditional metrics that often struggle to quantify reliability during testing. While repeatability shows comparable results between the CNN and CNN-BiLSTM models, the SUE metric provides critical insights into model reliability under noise (Fig. 7). The analyses stated the superiority of the CNN-BiLSTM model and revealed limitations in the robustness of DenseNet.

Furthermore, we assess the relationship between the SUE score and prediction accuracy during testing. The SUE metric provides insights into model prediction reliability beyond just repeatability. By comparing SUE values between correctly and incorrectly classified cases, we can evaluate model robustness to noise and other distortions. As depicted in Fig. 8, SUE values consistently showed higher values for correctly classified instances compared to misclassified ones. This consistent gap in SUE between accurate and inaccurate predictions demonstrates the metric's usefulness for assessing reliability.

The final analysis involved comparing the proposed SUE method with traditional uncertainty estimation methods, namely Deep Ensemble and MCD. The uncertainty values were evaluated for all four types of noise utilized in this study. While Deep Ensemble and SUE demonstrate similar performance in assessing uncertainty for correctly and misclassified samples in the test set, SUE offers several advantages. Deep Ensemble requires three trained models, resulting in increased computational costs in real-world scenarios. In contrast, SUE only requires a single model for uncertainty estimation. Furthermore, SUE exhibits excellent performance, particularly when dealing with ECG signals corrupted by real noise, such as motion and electrode artifacts.

Another noteworthy aspect is that SUE serves as a bridge between XAI and UQ. SUE offers the ability to extract additional information compared to Grad-CAM. Unlike Grad-CAM, SUE goes beyond evaluating feature importance within the original signal. It assesses the spatial overlap of the most critical features, offering quantitative insights into the network's robustness in terms of both current predictions and prediction reliability. Moreover, SUE is a highly versatile metric that can be applied to any network compatible with Grad-CAM, without requiring model retraining. This versatility allows for its seamless integration into existing models for enhanced interpretability and uncertainty assessment. In this work, we also presented the CNN-BiLSTM model, which outperformed both the CNN and DenseNet models.

In summary, the proposed methodology offers several key benefits:

1) SUE quantifies the spatial overlap of the most relevant features from Grad-CAM, providing a confidence score for the current prediction. We are the first group to propose SUE for CAD ECG signals.
2) SUE values have a strong correlation with classification accuracy during test time.
3) SUE can be easily applied during inference without the need for network retraining.
4) SUE can identify the most suitable network by evaluating Grad-CAM repeatability using synthesized ECG signals with diverse noise types and intensities.
5) SUE quantitative information on the network's robustness, encompassing both current prediction accuracy and the repeatability of predictions

It is important to recognize the limitations inherent in the proposed method. In this study, SUE is used specifically to assess the spatial overlap of features and the prediction's robustness. However, exploring the applications of SUE reveals numerous possibilities and diverse uses. Future investigations could focus on integrating SUE into UQ models to create an indicator that evaluates multiple aspects of predictions, including both uncertainty and repeatability. In this contest, an area of work could be the integration of SUE with probabilistic uncertainty estimation techniques such as Monte Carlo dropout or ensembling to provide a more comprehensive evaluation of model reliability from different perspectives. Indeed, quantifying both spatial and probabilistic uncertainty could offer additional insights into robustness.

Furthermore, future studies could examine the scalability of the model by applying SUE under real-world conditions. This could involve augmenting the dataset by integrating it with real-time ECG monitoring systems. The feasibility of implementing our proposed pipeline in real-time is supported by the low inference time, as the CNN-BiLSTM model demonstrates an inference time of only 0.05 s for a 2-second ECG segment. This aspect reaffirms the practicality of using our pipeline in real-time applications.

Moreover, there is potential for deeper exploration in future studies regarding the interpretability of the model and SUE from a

clinical perspective. This exploration could shed light on how physicians can leverage these insights in the decision-making process and evaluate the model's alignment with clinical criteria for CAD diagnosis. Future studies could also expand the application of this pipeline to various tasks, such as classifying other cardiac pathologies or different physiological signals. While this work focuses on validating SUE using ECG signals due to their importance in cardiac health applications, future studies could investigate applying SUE in the classification of neurological pathologies using electroencephalography (EEG) [51] and electromyography (EMG). Moreover, a promising avenue for future investigations could involve an in-depth exploration of the model's sustained performance and its capacity to adapt to dynamic patterns in ECG data. Delving into the implications of periodic retraining or fine-tuning with new data may yield valuable insights, contributing to the ongoing refinement of the model's long-term robustness.

6. Conclusion

In this study, we have introduced a novel metric called SUE, which integrates aspects of XAI with UQ principles, providing quantitative insights into the model's robustness. By assessing the spatial overlap of discriminative features, the SUE quantitatively evaluates network reliability on a scale from 0 to 1, encompassing both current prediction accuracy and the repeatability of predictions.

Initially, we proposed a CNN-BiLSTM deep learning network for accurate ECG signal classification in CAD, surpassing state-of-the-art methods and representing a significant advancement. To validate the SUE, we conducted a comprehensive analysis under four noise conditions applied to the ECG signal, involving two synthetic noises (random noise and power line interference) and two real noises (motion and electrode movement artifacts). The SUE not only highlights the CNN-BiLSTM model's superior performance in noisy conditions but consistently identifies the top-performing network, elucidating limitations within CNN and DenseNet. Additionally, we observed a clear correlation between the SUE and prediction accuracy, with higher SUE values associated with correctly classified cases and lower values with misclassified instances. Finally, the SUE showcases comparable performance to traditional uncertainty estimation methods, Deep Ensemble and Monte Carlo Dropout, when evaluating uncertainty for correctly and misclassified samples within the test set. In summary, this research not only introduces a novel reliability metric for AI models but also underscores the potential of combining XAI and UQ techniques to enhance ECG analysis. The proposed method can be employed for other healthcare applications using physiological signals such as phonocardiogram (PCG), EEG, photoplethysmography (PPG), EMG, etc.

CRediT authorship contribution statement

Silvia Seoni: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Filippo Molinari:** Writing – review & editing, Supervision, Methodology, Conceptualization. **U. Rajendra Acharya:** Writing – review & editing, Supervision, Methodology. **Oh Shu Lih:** . **Prabal Datta Barua:** Writing – review & editing, Supervision. **Salvador García:** Writing – review & editing, Supervision, Methodology. **Massimo Salvi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this work are publicly available

References

[1] World Health Organization, Global status report on noncommunicable diseases 2014, World Health Organization, 2014.
[2] L. Maximilian Buja, J.T. Willerson, The role of coronary artery lesions in ischemic heart disease: insights from recent clinicopathologic, coronary arteriographic, and experimental studies, Hum. Pathol. 18 (5) (1987) 451–461, https://doi.org/10.1016/S0046-8177(87)80030-8.
[3] L. M. Buja and H. A. McAllister, "Atherosclerosis: Pathologic Anatomy and Pathogenesis," Cardiovasc. Med., pp. 1581–1591, 2007, doi: 10.1007/978-1-84628-715-2_76.
[4] N. Herring and D. J. Paterson, "ECG diagnosis of acute ischaemia and infarction: past, present and future," QJM: An International Journal of Medicine, vol. 99, no. 4, pp. 219–230, Apr. 2006, doi: 10.1093/QJMED/HCL025.
[5] F. Fein, "Heart disease in diabetes mellitus: theory and practice.," Diabetes mellitus: theory and practice, pp. 812–823, 1990.
[6] R.J. Martis, U.R. Acharya, H. Adeli, Current methods in electrocardiogram characterization, Comput. Biol. Med. 48 (1) (2014) 133–149, https://doi.org/10.1016/J.COMPBIOMED.2014.02.012.
[7] R. Aggarwal, P. Podder, A. Khamparia, ECG classification and analysis for heart disease prediction using XAI-driven machine learning algorithms, Intelligent Systems Reference Library 222 (2022) 91–103, https://doi.org/10.1007/978-981-19-1476-8_7/COVER.
[8] S. Matin Malakouti, "Heart disease classification based on ECG using machine learning models," Biomed Signal Process Control, vol. 84, p. 104796, Jul. 2023, doi: 10.1016/J.BSPC.2023.104796.
[9] S. S. Al-Zaiti et al., "Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction," Nature Medicine 2023 29:7, vol. 29, no. 7, pp. 1804–1813, Jun. 2023, doi: 10.1038/s41591-023-02396-3.
[10] M. Hassaballah, Y. M. Wazery, I. E. Ibrahim, and A. Farag, "ECG Heartbeat Classification Using Machine Learning and Metaheuristic Optimization for Smart Healthcare Systems," Bioengineering 2023, Vol. 10, Page 429, vol. 10, no. 4, p. 429, Mar. 2023, doi: 10.3390/BIOENGINEERING10040429.

[11] U. R. Acharya et al., "Entropies for automated detection of coronary artery disease using ECG signals: A review," Biocybernetics and Biomedical Engineering, vol. 38, no. 2. PWN-Polish Scientific Publishers, pp. 373–384, Jan. 01, 2018. doi: 10.1016/j.bbe.2018.03.001.
[12] A. Asgharzadeh-Bonab, M. Chehel Amirani, and A. Mehri, "Spectral entropy and deep convolutional neural network for ECG beat classification," 2020, doi: 10.1016/j.bbe.2020.02.004.
[13] V. Jahmunah, E.Y.K. Ng, T.R. San, U. Rajendra Acharya, Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals, Comput. Biol. Med. 134 (2021) 104457, https://doi.org/10.1016/j.compbiomed.2021.104457.
[14] V. Jahmunah, E.Y.K. Ng, R.-S. Tan, S. Lih Oh, U. Rajendra Acharya, Uncertainty quantification in DenseNet model using myocardial infarction ECG signals, Comput. Methods Programs Biomed. 229 (2023) 107308, https://doi.org/10.1016/j.cmpb.2022.107308.
[15] Y. Xu, S. Zhang, W. Xiao, Inter-patient ECG classification with intra-class coherence based weighted kernel extreme learning machine, Expert Syst. Appl. 227 (Oct. 2023) 120095, https://doi.org/10.1016/J.ESWA.2023.120095.
[16] H. Mewada, 2D-wavelet encoded deep CNN for image-based ECG classification, Multimed. Tools Appl. 82 (2023) 20553–20569, https://doi.org/10.1007/s11042-022-14302-z.
[17] H.M. Rai, K. Chatterjee, Hybrid CNN-LSTM deep learning model and ensemble technique for automatic detection of myocardial infarction using big ECG data, Appl. Intell. 52 (2022) 5366–5384, https://doi.org/10.1007/s10489-021-02696-6/Published.
[18] N. Razmjooy, A. Arshaghi, Application of multilevel thresholding and CNN for the diagnosis of skin cancer utilizing a multi-agent fuzzy buzzard algorithm, Biomed. Signal Process. Control 84 (Jul. 2023), https://doi.org/10.1016/j.bspc.2023.104984.
[19] Q. Huang, H. Ding, N. Razmjooy, Optimal deep learning neural network using ISSA for diagnosing the oral cancer, Biomed. Signal Process. Control 84 (Jul. 2023), https://doi.org/10.1016/j.bspc.2023.104749.
[20] F. Bayram, A. Eleyan, COVID-19 detection on chest radiographs using feature fusion based deep learning, Signal Image Video Process 16 (6) (Sep. 2022) 1455–1462, https://doi.org/10.1007/s11760-021-02098-8.
[21] C. Yan, N. Razmjooy, Kidney stone detection using an optimized deep believe network by fractional coronavirus herd immunity optimizer, Biomed. Signal Process. Control 86 (Sep. 2023), https://doi.org/10.1016/j.bspc.2023.104951.
[22] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," Computer Methods and Programs in Biomedicine, vol. 226. Elsevier Ireland Ltd, Nov. 01, 2022. doi: 10.1016/j.cmpb.2022.107161.
[23] R.R. Selvaraju, et al., Grad-CAM: visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2020) 336–359, https://doi.org/10.1007/s11263-019-01228-7.
[24] S. Seoni, et al., Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023), Comput. Biol. Med. (2023), https://doi.org/10.1016/j.compbiomed.2023.107441.
[25] Y.Y. Jo, et al., Detection and classification of arrhythmia using an explainable deep learning model, J. Electrocardiol. 67 (2021) 124–132, https://doi.org/10.1016/j.jelectrocard.2021.06.006.
[26] J.W. Hughes, et al., Performance of a convolutional neural network and explainability technique for 12-Lead electrocardiogram interpretation, JAMA Cardiol. 6 (11) (2021) 1285–1295, https://doi.org/10.1001/jamacardio.2021.2746.
[27] B.M. Maweu, S. Dakshit, R. Shamsuddin, B. Prabhakaran, CEFEs: a CNN explainable framework for ECG signals, Artif. Intell. Med. 115 (2021) 102059, https://doi.org/10.1016/J.ARTMED.2021.102059.
[28] R.R. Van De Leur, et al., Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders, European Heart Journal - Digital Health 3 (2022) 390–404, https://doi.org/10.1093/ehjdh/ztac038.
[29] R. Varandas, B. Gonçalves, H. Gamboa, P. Vieira, Quantified explainability: convolutional neural network focus assessment in arrhythmia detection, BioMedInformatics 2 (1) (2022) 124–138, https://doi.org/10.3390/biomedinformatics2010008.
[30] J.Y. Park, K. Lee, N. Park, S.C. You, J.G. Ko, Self-attention LSTM-FCN model for arrhythmia classification and uncertainty assessment, Artif. Intell. Med. 142 (2023) 102570, https://doi.org/10.1016/J.ARTMED.2023.102570.
[31] Y. Elul, A. A. Rosenberg, A. Schuster, A. M. Bronstein, and Y. Yaniv, "Meeting the unmet needs of clinicians from AI systems showcased for cardiology with deep-learning-based ECG analysis MEDICAL SCIENCES," vol. 118, p. 2020620118, 2021, doi: 10.1073/pnas.2020620118/-/DCSupplemental.
[32] M. Barandas, et al., Evaluation of uncertainty quantification methods in multi-label classification: a case study with automatic diagnosis of electrocardiogram, Information Fusion 101 (2024), https://doi.org/10.1016/j.inffus.2023.101978.
[33] J.F. Vranken, et al., Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms, European Heart Journal - Digital Health 2 (3) (2021) 401–415, https://doi.org/10.1093/ehjdh/ztab045.
[34] D. Zhang, S. Yang, X. Yuan, and P. Zhang, "Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram," iScience, vol. 24, no. 4, p. 102373, Apr. 2021, doi: 10.1016/J.ISCI.2021.102373.
[35] J. Belen, S. Mousavi, A. Shamsoshoara, and F. Afghah, "An Uncertainty Estimation Framework for Risk Assessment in Deep Learning-based Atrial Fibrillation Classification," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2011.00121.
[36] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet Components of a New Research Resource for Complex Physiologic Signals," 2000. [Online]. Available: http://www.physionet.org.
[37] B.N. Singh, A.K. Tiwari, Optimal selection of wavelet basis function applied to ECG signal denoising, Digit Signal Process 16 (3) (May 2006) 275–287, https://doi.org/10.1016/J.DSP.2005.12.003.
[38] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681, https://doi.org/10.1109/78.650093.
[39] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (Nov. 1997) 1735–1780, https://doi.org/10.1162/NECO.1997.9.8.1735.
[40] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw. 5 (2) (1994) 157–166, https://doi.org/10.1109/72.279181.
[41] U.R. Acharya, H. Fujita, O.S. Lih, M. Adam, J.H. Tan, C.K. Chua, Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network, Knowl Based Syst 132 (2017) 62–71, https://doi.org/10.1016/j.knosys.2017.06.003.
[42] M. D'Aloia, A. Longo, and M. Rizzi, "Noisy ECG Signal Analysis for Automatic Peak Detection," Information 2019, Vol. 10, Page 35, vol. 10, no. 2, p. 35, Jan. 2019, doi: 10.3390/INFO10020035.
[43] K.L. Venkatachalam, J.E. Herbrandson, S.J. Asirvatham, Signals and signal processing for the electrophysiologist part I: electrogram acquisition, Circ. Arrhythm. Electrophysiol. (2011), https://doi.org/10.1161/CIRCEP.111.964304.
[44] R.G.M. George, B. Moody, and and W.K. Muldrow, "A NOISE STRESS TEST," for Arrhythmia Detectors, pp. 381–384, 1984.
[45] Y. Gal and Z. Ghahramani, "Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference," Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.02158.