IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 February 2024, accepted 27 March 2024, date of publication 3 April 2024, date of current version 18 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3384277

RESEARCH ARTICLE

Sig-Lime: A Signal-Based Enhancement of Lime Explanation Technique

TALAL ALI AHMED ABDULLAH[1], MOHD SOPERI MOHD ZAHID[1], (Member, IEEE), AHMAD F. TURKI[2,3], WALEED ALI[4], AHMAD A. JIMAN[2,3], MOHAMMED J. ABDULAAL[2,3], NEBRAS M. SOBAHI[2], AND EYAD T. ATTAR[2]

[1]Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Malaysia
[2]Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[3]Center of Excellence in Intelligent Engineering Systems, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[4]Information Technology Department, Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah 25729, Saudi Arabia

Corresponding author: Mohd Soperi Mohd Zahid (msoperi.mzahid@utp.edu.my)

This work was supported by the Institutional Fund Project provided by the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia, under Grant IFPIP: 1313-135-1443.

ABSTRACT Interpreting machine learning models is facilitated by the widely employed locally interpretable model-agnostic explanation (LIME) technique. However, when extending LIME to signal data, its credibility falters due to perturbation techniques used to generate local datasets. These techniques disrupt temporal dependencies among features, leading to unrealistic data points and potentially misleading explanations. Additionally, LIME faces instability and local fidelity issues, limiting its suitability for real-world applications. The absence of a dedicated LIME package tailored for interpreting signal data further diminishes comprehensibility, especially when applied to models trained on such data. In this paper, we introduce Signal-based LIME (Sig-LIME) to address these limitations. Sig-LIME leverages a novel data generation technique that captures temporal dependence among features, enhancing credibility and stability. It combines a random forest model and heatmaps to provide illuminating explanations for predictions drawn from electrocardiogram (ECG) signals, improving model transparency. Empirical findings underscore the enhanced interpretability and comprehension of model predictions attained by Sig-LIME compared to baseline LIME. Our quantitative evaluation based on an analysis of variance (ANOVA) framework, reveals a notable improvement in stability with Sig-LIME, evidenced by an f-statistic of 0.0 and p-values of 1, indicating a complete absence of variation between multiple runs. Regarding local fidelity, Sig-LIME surpasses the baseline LIME, exhibiting a lower average Euclidean distance of 0.49 compared to 17.24. Sig-LIME excels in generating data more akin to the original, achieving remarkable stability and significantly enhancing credibility and local fidelity in the explanations it generates.

INDEX TERMS Deep learning, ECG signal, explanation, LIME, Sig-LIME, cardiac arrhythmia, artificial intelligence.

I. INTRODUCTION

Interpretability is a crucial aspect in the field of machine learning (ML), enabling the understanding and explanation of decisions made by ML models in a coherent and meaningful way [1], [2]. Although a precise mathematical definition of interpretability remains elusive, researchers have highlighted its significance through various perspectives and definitions [3]. Among these perspectives, a commonly accepted concept focuses on the extent to which individuals can understand the rationale behind a machine learning model's decision [4]. This perspective emphasizes the importance of human comprehension and the ability to identify the causal factors that influence the model's predictions.

The importance of interpretability in deep learning is especially prominent in critical domains, such as healthcare [5],

The associate editor coordinating the review of this manuscript and approving it for publication was Frederico Guimarães.

and security [6], [7]. In these domains, the reliability of decision-making processes holds significant implications, especially as legal frameworks in certain regions [8] require medical professionals to provide justifications for their diagnoses. However, the lack of interpretability in deep learning models poses a significant obstacle to their real-world applicability [9]. Therefore, unraveling the decision-making processes within these models is essential to ensure not only the accuracy, fairness, and ethical nature of predictions but also to cultivate a higher degree of trust in situations where decisions have a substantial impact on individuals' lives [10].

Improving the understandability of complex deep learning models has led to the emergence of the field of interpretable deep learning. Techniques encompass a wide range, including generating graphical interpretations, defining IF-THEN rules, and assigning weights to features, all with the goal of clarifying the opaque logic of complex models [11]. This endeavor has resulted in numerous innovative approaches that can be categorized based on factors such as intrinsic or post-hoc complexity, model-agnostic or model-specific nature, and local or global scope. Notably, model-agnostic techniques have received significant attention due to their capability to interpret the decision-making process of any machine learning model without modifying its internal workings, as extensively discussed in [10]. The majority of contemporary research works in the field of Explainable Artificial Intelligence (XAI) has concentrated on devising methodologies primarily for computer vision and natural language processing applications while the exploration and development of methods for time series data have received comparatively less [12].

Among these techniques, the locally interpretable model-agnostic explanation (LIME) method [13] serves as an example of a model-agnostic approach. LIME plays a crucial role in approximating the accuracy of specific ML models. This method entails perturbation strategies that create surrogate data points for making predictions. Subsequently, an interpretable ML model is trained to explain the underlying reasons for these predictions [14]. However, it is important to note a significant limitation of LIME. In the baseline LIME's perturbation technique disrupts the temporal relationships present in signals, such as ECG and EEG [15], potentially resulting in unlikely and unrealistic data points [16].

Furthermore, LIME faces challenges like instability and local fidelity, reducing its suitability for real-world applications [17], [18], [19], [20]. The randomness introduced by LIME's perturbation technique results in different samples and explanations for the same instance, rendering the explanations inconsistent and unreliable. The challenge of local fidelity arises from the complex temporal dependencies among attributes in time series data, which are essential for generating valid and substantial explanations. Independently perturbing data disrupts these temporal relationships, leading to explanations that do not adequately represent the local behavior of the deep learning model. Furthermore, the absence of a dedicated LIME package tailored for interpreting signal data results in less comprehensible outcomes, especially when applied to models trained on such data [16].

This work introduces the Signal-based LIME (Sig-LIME) technique to address the above-mentioned limitations of LIME. The main focus is to enhance the interpretability of deep learning models in the context of signal data, with particular emphasis on ECG signals.

Sig-LIME employs a novel data generation approach to provide explanations at the signal level for models that work with signals such as electrocardiogram (ECG) signals. This method maintains the temporal correlations between features through signal segmentation and controlled noise introduction. By utilizing this strategy, Sig-LIME enhances decision-making and extracts more valuable insights from complex signal datasets.

To provide comprehensive explanations for predictions derived from ECG signals, Sig-LIME also integrates a Random Forest model and heatmaps. The Random Forest model, capable of capturing complex feature relationships in the data, enhances the depth and accuracy of explanations. Heatmaps offer a visually intuitive representation of feature contributions to model predictions, making it easier to understand complex model behaviors. The primary contributions of this work can be summarized as follows:

- Advancing Interpretability for Deep Learning Models in Signal Data: This study introduces the Signal-based LIME (Sig-LIME) method, enhancing the interpretability of deep learning models in the context of signal data analysis, with a particular focus on applications such as electrocardiogram (ECG) signals. Sig-LIME provides a framework for effectively unraveling the complex decisions made by these models.
- Mitigation of LIME Limitations in Signal Data: To overcome the constraints of the locally interpretable model-agnostic explanation (LIME) technique, Sig-LIME introduces innovative data generation techniques and incorporates a random forest model and heatmaps. By preserving temporal relationships and addressing issues of instability and local fidelity, Sig-LIME offers more reliable and credible explanations for interpreting signal data.
- Enhanced Precision in Model Explanations: Sig-LIME ensures the precision and comprehensibility of explanations by preserving temporal feature relationships in signal data. Through controlled noise and signal segmentation techniques, it enhances the decision-making process, leading to valuable insights from complex datasets. Furthermore, the combination of a random forest model and heatmaps provides transparent and intuitive explanations for ECG signal predictions.

The paper is organized as follows: It commences by reviewing related works in interpretable machine learning in Section II. Subsequently, it delves into the limitations of LIME explanations, particularly in the context of signal data, in Section III. Section IV introduces Sig-LIME, an enhanced

approach. Section V presents experimental results showcasing Sig-LIME's effectiveness in enhancing interpretability for a hybrid CNN-GRU model. Finally, Section VI provides the conclusion, summarizing the findings, emphasizing contributions, and suggesting future research directions.

II. RELATED WORKS

In recent years, machine learning (ML) and deep learning (DL) have emerged as transformative forces in classifying cardiac arrhythmias. For instance, [21] showcases the innovative use of convolutional neural networks (CNNs) in arrhythmia detection, leveraging advanced optimization techniques to enhance model performance. This signifies a substantial leap in accuracy and efficiency for ECG analysis. Similarly, [22] highlights the nuanced capabilities of ML algorithms in differentiating between various ECG beat types.

The rapid advancements in ML and DL techniques highlight their transformative impact on cardiac healthcare diagnostics. Yet, the role of interpretability in machine learning remains essential, as it ensures clear and meaningful understanding of model decisions [1], [2]. Concurrently, the field of interpretable machine learning has seen significant growth in methods enhancing model transparency [23], [24], [25], crucial for building trust and comprehensibility in healthcare AI applications.

Among these approaches, the Locally Interpretable Model-Agnostic Explanation (LIME) technique stands out as a prominent method for generating local explanations to elucidate individual predictions [12]. By approximating the behavior of intricate models, LIME addresses the challenge of interpretability by creating surrogate models that capture the model's behavior around a specific instance of interest [26].

However, researchers have identified certain limitations within LIME that have prompted them to explore extensions and modifications in order to address these challenges. One such extension is Guided-LIME [14], which introduces a hybrid approach by integrating LIME with Formal Concept Analysis (FCA) for structured sampling of instances to enhance LIME global explanation. OptiLIME [27] addresses the trade-off between the LIME explanation's stability and fidelity and proposed a framework that automatically finds the best kernel width that maximizes the stability while retaining a predefined level of fidelity.

ALIME [20] on the other hand, employs an autoencoder as a weighting function for the local model to improve robustness and accuracy of the local model. It achieves this by adding a small amount of white Gaussian noise to the training data and utilizing the autoencoder to compute the latent representations for both the explanation instance and the generated points, which are subsequently weighted using an exponential kernel as a distance measure.

MeLIME [28] proposes three key modifications to LIME: data generation using different kernels, improvements to the explanation generation method, and adaptive sample size adjustment based on data complexity. These enhancements aim to enhance the accuracy, interpretability, and adaptability of LIME's explanations for better understanding of machine learning models. K-LIME [29] utilizes an unsupervised learning technique called k-means to partition the dataset into K clusters. Each cluster then applies a local general linear model, and the K is adjusted to maximize the R2 for all local models.

LIME-SUP [30] replaces the unsupervised clustering technique used in K-LIME with a supervised partitioning tree to enhance the quality of explanations. NormLIME [31] introduces a class-specific global explanation by aggregating and normalizing a group of defined local explanations. DLIME [32] proposed a deterministic version of LIME, replacing the random perturbation technique with Agglomerative Hierarchical Clustering to cluster the data and employing K-nearest Neighbor to select the relevant cluster for the instance being explained.

LIME-Aleph [33] applies the Inductive Logic Programming approach Aleph to generate explanations in the form of logic rules that capture combinations of features and their relationships. audioLIME [11] leverages source separation algorithms to provide audibly interpretable explanations based on waveform predictions. The perturbations used in audioLIME are created by switching on/off components extracted by source separation to generate listenable explanations.

While various methods have been proposed to improve LIME's instability and lack of fidelity in general, they might not be directly suitable for signal data like ECG signals, given their unique characteristics and temporal dependencies [34]. The temporal dependencies and unique characteristics of signal data further exacerbate the issues associated with the random perturbation technique [20].

To address these challenges, it is crucial to create specialized enhancements for LIME when it is applied to signal data. Complementing this discourse, our previous study B-LIME [35] presents a notable advancement in model interpretability. B-LIME improves upon the original LIME framework, tailoring it for the specific nuances of ECG signal analysis. This adaptation signifies a crucial step towards making deep learning models in cardiac healthcare more transparent and comprehensible. Our work with Sig-LIME aligns closely with this objective, aiming to further enhance the interpretability of DL models in signal data analysis.

The Sig-LIME method introduces enhancements to the data generation and explanation techniques used by LIME, designed specifically for generating signal-level explanations for models that take signals as input, such as ECG signals. Sig-LIME considers the unique characteristics of the data, preserving temporal dependencies and ensuring local fidelity. As a result, the explanations produced by the Sig-LIME method are more accurate, credible, and reliable, enabling healthcare professionals to gain deeper insights into the model's decisions and enhance clinical decision-making.

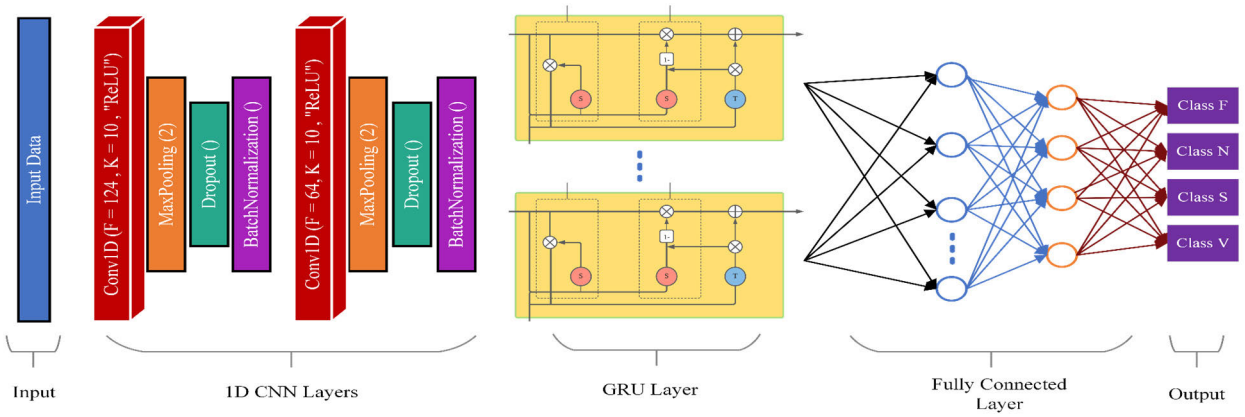

FIGURE 1. The framework of the hybrid model. The model combines two blocks of 1D CNN and one block of GRU structure. Each 1D CNN block contains a max-pooling, a dropout, a batch-normalization layer.

III. EXPLORING LIME EXPLANATIONS: INVESTIGATING OUTCOMES AND LIMITATIONS

With a focus on electrocardiogram (ECG) signals, we comprehensively explore the application of local interpretation through model-agnostic techniques, specifically LIME, within the context of signal data in this section. These investigations encompass the development of deep learning models, including gated recurrent unit (GRU) algorithms and one-dimensional convolutional neural networks (1D CNN). The primary goal is to utilize LIME to furnish explanations for the predictions made by these hybrid models, with a specific emphasis on evaluating the interpretability, reliability, stability, and local validity of each explanation.

A. HYBRID DEEP LEARNING MODEL FOR CARDIAC ARRHYTHMIAS CLASSIFICATION

To begin our investigation, we built a hybrid deep learning model that effectively combines the characteristics of 1D CNN and GRU algorithms. The goal is to categorize ECG data into four different categories of cardiac arrhythmias. For the purpose of accurately capturing spatial and temporal relationships in the ECG data, the proposed model architecture combines 1D CNN and GRU. [36]. This hybrid design capitalizes on the characteristics of both techniques, enabling effective feature extraction from signals while also capturing temporal relationships, hence improving arrhythmia classification accuracy [37], [38].

The 1D CNN component serves as a feature extractor [39], [40], while the GRU component acts as a sequence modeler enabling the model to provide a posterior probability for the presence of arrhythmias in the input sequence, leading to higher classification accuracy [37]. The model architecture is illustrated in Fig. 1.

The 1D CNN component comprises of two convolutional layers, each followed by a max-pooling, dropout, and batch-normalization layer, as well as a rectified linear unit (ReLU) activation function. This configuration enables the model to absorb hierarchical information from the input ECG data. The convolutional layers process the input signals using a collection of teachable filters to extract regional information important for classifying arrhythmias. By introducing non-linearity, the ReLU activation function allows the model to recognize complicated patterns [41]. The max-pooling layers reduce the spatial dimensionality while maintaining the most important characteristics by down-sampling the feature maps [42]. The dropout layer mitigates overfitting by randomly setting inputs to zero during training with a specified rate [43]. This helps in preventing the network from relying too heavily on specific features. Furthermore, the batch-normalization layer normalizes inputs within each mini-batch, leading to a more efficient and stable learning process [44].

The GRU component is then given the output of the 1D CNN component. GRU is a form of recurrent neural network (RNN) that is particularly successful at representing sequential data [45]. By maintaining a concealed state that integrates data from earlier time steps, it can capture long-term dependencies. The model's GRU layer enables the modeling of temporal dependencies in the ECG data, allowing it to recognize patterns that change over time [46].

The fully connected layer in the proposed deep learning model consists of two dense layers. The first dense layer has 512 nodes with a rectified linear unit (ReLU) [47] activation function, introducing non-linearity to the network. The second dense layer has four nodes representing the four classes of arrhythmias and uses the softmax activation function. This function provides class probabilities, which show the chance that each arrhythmia class the supplied ECG data belongs to [48]. The softmax activation layer enables interpretation of the model's level of confidence in its predictions, with the class with the highest probability being taken into account as the anticipated arrhythmia class.

Backpropagation and gradient descent optimization methods are used to train the model. Sparse cross-entropy is the employed loss function, which assesses the dissimilarity between the anticipated probability and the actual labels [49].

During training, the model learns the optimal weights and biases that minimize the loss function and improve the accuracy of the predictions.

B. UTILIZING LIME FOR EXPLANATIONS

LIME serves as our chosen approach for generating explanations to shed light on the predictions made by the hybrid CNN-GRU model. By perturbing instances of interest and observing the model's responses, LIME approximates the local behavior of the complex model, thus providing interpretable insights into the decision-making process [14]. In this exploration, LIME is applied to elucidate the hybrid model's predictions for cardiac arrhythmias, with a focus on enhancing understandability and transparency.

To facilitate the integration of LIME with our Hybrid 1D CNN-GRU model, we leverage the "RecurrentTabularExplainer" package, specifically tailored for handling multidimensional inputs, such as sequential data processed by deep learning models [50], [51]. As shown in Fig. 2, the integration procedure involves the model and the instance (the sample being explained).

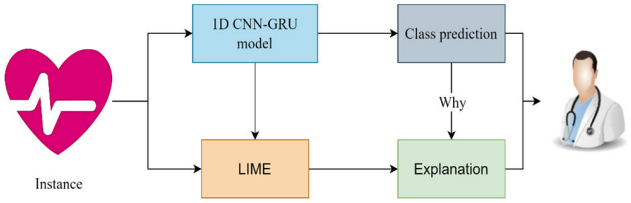

FIGURE 2. Integrating LIME procedure. LIME requires a sample (heartbeat) to explain and a DL model (1D CNN-GRU) to generate explanation.

The LIME explanation process begins by selecting a specific heartbeat from the MIB-BIH dataset that requires interpretation. This instance is presented to the Hybrid CNN-GRU model, and its prediction is obtained. LIME then perturbs the features of the instance while keeping the label constant, generating a dataset of perturbed instances. These instances are used to train an interpretable surrogate model, such as Ridge regression model, that approximates the behavior of the Hybrid CNN-GRU model in the vicinity of the selected instance.

RecurrentTabularExplainer is a specific module of LIME designed to handle multidimensional input, particularly sequential data processed by deep learning models [52]. The result of the LIME integration is presented in Fig. 3, which encapsulates the explanations meticulously generated by this method. This figure features three insightful plots, each offering a unique perspective into the underlying mechanisms of the Hybrid CNN-GRU model's decision-making process.

The prediction probabilities plot, exemplified in the figure, portrays the distribution of predicted probabilities across diverse classes. It provides insights into how the model allocates probabilities to various potential outcomes based on the input features.

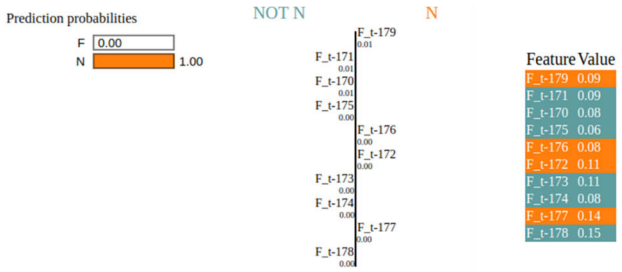

FIGURE 3. LIME explanation. Due to the absence of a dedicated backage for signal data, researchers usually use the tabular data package which mainly focuses on the feature names.

Furthermore, the feature importance plot, a key element of the ensemble, presents a ranked exposition of individual feature significances in steering the model's predictions. Each feature carries a distinct weight indicative of its influence on the model's output. This plot empowers users to discern pivotal features, identifying those that exert maximum impact on predictions, ultimately revealing the fundamental rationale guiding the model's decisions.

Lastly, the feature value plot takes center stage, depicting the intricate interplay between individual feature values and the model's resultant output. This plot highlights the delicate relationship between specific feature values and their corresponding influences on predictions. It offers a clear understanding of both the direction and magnitude of these effects, demystifying the complex associations between feature attributes and predictions. Another feature value plot, elegantly displayed in Fig. 4, further enriches this narrative.

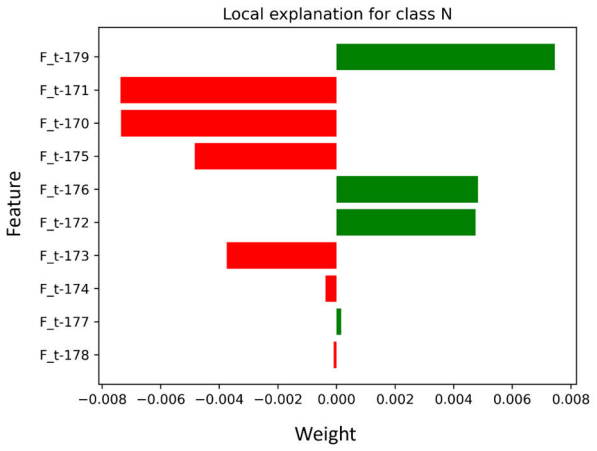

FIGURE 4. Feature value plot of LIME explanation. The x axis represents the feature weight while the y axis represents the feature name.

C. INVESTIGATING LIME EXPLANATIONS ON THE HYBRID 1D CNN-GRU

In this section, we embark on a profound exploration of the application of LIME approach within the context of the Hybrid 1D CNN-GRU model for cardiac arrhythmia classification.

Our investigation is guided by a comprehensive assessment of LIME explanations based on the interpretation properties elucidated in [10], which encompass Understandability, Fidelity, Stability (Robustness), and local fidelity, all of which are crucial elements in establishing the reliability and comprehensibility of model explanations.

Guiding our investigative expedition are the following central inquiries that underpin our understanding and evaluation of LIME explanations:

- Understandability: We delve into the extent to which the provided explanations are inherently understandable, catering to diverse audiences, including medical practitioners and domain experts.
- Fidelity and Credibility: We scrutinize the credibility of the explanations, assessing the extent to which they can be relied upon for their accuracy and reliability.
- Stability: We investigate the consistency and stability of LIME explanations across diverse scenarios, encompassing different runs and perturbations.
- Local Fidelity: Recognizing LIME's inherently local focus, we inquire into whether the explanations faithfully and accurately capture the behavior of the Hybrid 1D CNN-GRU model within the specific context of individual instances.

As depicted in Fig. 5, these pivotal questions act as the compass that guides our inquiry, steering us toward a comprehensive understanding of the symbiotic relationship between LIME explanations and the intricate landscape of cardiac arrhythmia classification. Through our meticulous investigation, we aim not only to shed light on the nuances of model interpretability but also to ascertain the potential of LIME to augment healthcare practices by offering transparent and elucidated predictive insights.

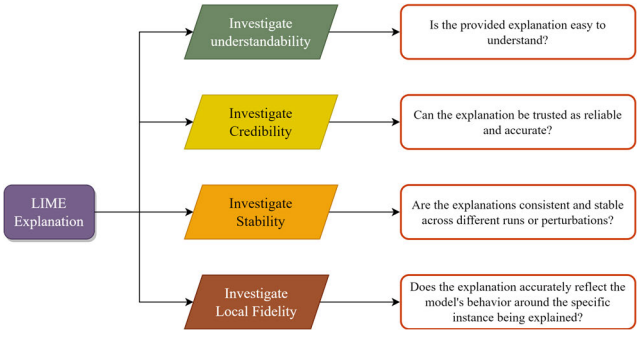

FIGURE 5. Guideline of LIME investigation. The main aim is to investigate LIME explanation based on main interpretation properties.

1) INVESTIGATING THE UNDERSTANDABILITY OF LIME EXPLANATIONS

Explanation understandability is a crucial aspect when using LIME method to interpret machine learning models, particularly in the context of signal data such as ECG signals.

However, a significant challenge arises from the absence of a dedicated package in LIME that is specifically tailored for signal data interpretation. This limitation contributes to the provision of less understandable, especially when applied to models trained on signal data.

Signal data, including ECG signals, possess unique characteristics and temporal dependencies that demand specialized treatment during the explanation process [53]. The traditional approach of LIME, which is designed primarily for tabular data, may not fully capture the intricate patterns and dynamics present in signal data [51]. This limitation becomes evident from the outcomes depicted in Fig. 3 and Fig. 4. Consequently, the explanations generated by LIME for signal data might not be as intuitive or easy to comprehend, especially for domain experts and medical practitioners who rely on clear and concise insights to make critical decisions.

One of the central issues arises from the nature of signal data, where feature names are represented as sequences of numbers, in our case from 0 to 180. This is different from tabular data, where each feature has a distinct label. As a result, when LIME attempts to explain signal features, the explanations can be ambiguous and unclear. This ambiguity stems from the absence of distinct feature names, making it challenging to translate LIME's insights into coherent and understandable information.

In the subsequent phases of our research, we will explore the heatmaps technique to improve the understandability of LIME explanations for signal data.

2) INVESTIGATING CREDIBILITY

The credibility of the explanations is of paramount importance in the integration of LIME with the Hybrid 1D CNN-GRU model. To ensure the trustworthiness and reliability of the insights provided by LIME, we investigate the model-agnostic nature of LIME's approach.

One of the primary concerns affecting the credibility of LIME explanations is the data perturbation technique used to generate local explanations. LIME employs perturbation by sampling data points in the vicinity of the instance being explained and fitting a linear model on these perturbed samples. While this approach is effective for tabular data, it may lead to unrealistic data points when dealing with complex signal data, such as ECG signals.

In Fig. 6, several samples of the data points generated by LIME are presented, providing visual evidence of the potential discrepancies between the generated data points and the actual signal. This is due to the distinctive temporal association between attributes in time series data, which must be considered to produce valid and substantial explanations. Signal data exhibits intricate temporal dependencies and unique characteristics, making it challenging to perturb the data effectively without introducing unrealistic values. The perturbation process may produce data points that do not align with the natural distribution of the original signal, leading to explanations that might not accurately reflect the model's behavior around the specific instance.

Another substantial limitation impacting the credibility of LIME explanations stems from the fundamental linearity

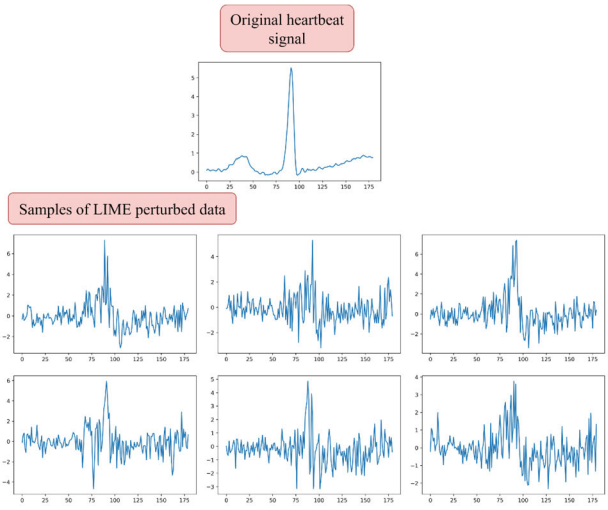

FIGURE 6. Guideline of LIME investigation. The main aim is to investigate LIME explanation based on main interpretation properties.

assumption inherent in the methodology [30]. LIME operates under the premise that the behavior of a machine learning model can be approximated by a locally linear model. However, this assumption might not hold true for intricate, non-linear deep learning models.

Deep learning models possess the capability to capture complex and non-linear relationships within data, enabling them to comprehend intricate patterns and representations. Nonetheless, attempting to approximate this non-linear behavior using a linear model during the explanation process can result in oversimplified and constrained interpretations. Such interpretations may fail to encompass the intricate complexity characterizing the decision-making process of deep learning models.

To quantify the reliability of LIME's surrogate model approximation for a specific instance, the method provides a score function. This score gauges how effectively the locally interpretable model mirrors the behavior of the underlying black-box model for that instance. It is instrumental in assessing the dependability of a local interpretation for a particular instance. A score approaching 1 indicates a faithful approximation of the black-box model's behavior in the vicinity of that instance. Conversely, a significantly lower score implies that the interpretable model might not be a reliable approximation.

For example, in our provided instance, LIME yields a score of 0.0008, which is remarkably close to 0. This outcome strongly aligns with our hypothesis, underscoring the limitations introduced by the linearity assumption in scenarios involving non-linear deep learning models.

3) INVESTIGATING STABILITY

Stability is another critical aspect to consider in the investigation. Instability refers to the variation or inconsistency in the explanations generated by LIME when the same instance is perturbed or when small changes are made to the data [54]. This inconsistency can undermine the reliability and trustworthiness of the explanations, hindering their effective utilization in decision-making processes. We assess whether the explanations generated by LIME exhibit consistency and stability across different runs or perturbations.

The data perturbation approach is one of the main causes of the instability in LIME. LIME creates a local dataset for fitting the interpretable model by randomly perturbing the original instance features. When the procedure is repeated, this random disturbance, however, might produce different samples, leading to several models and explanations for the same instance [35]. The explanations thus become vulnerable to the particular disturbances, making them untrustworthy and challenging to comprehend consistently.

The linearity assumption and feature selection procedure utilized in LIME are additional factors that contribute to explanation instability. LIME applies a linear model to the local dataset based on the assumption of linearity, choosing a subset of features that are thought to be crucial for the explanation. However, the feature selection process might be unstable, and change based on random perturbations, resulting in various sets of chosen features and explanations for the same occurrence.

Additionally, the local linear model might not be able to faithfully replicate the behavior of the complicated deep learning model if the linearity requirement is violated. As the linear model tries to reflect the complex non-linear relationships within the data, this restriction may make the explanations more unstable. An examination of the variations in LIME explanation outputs for the same instance over many runs is shown in Fig 7.

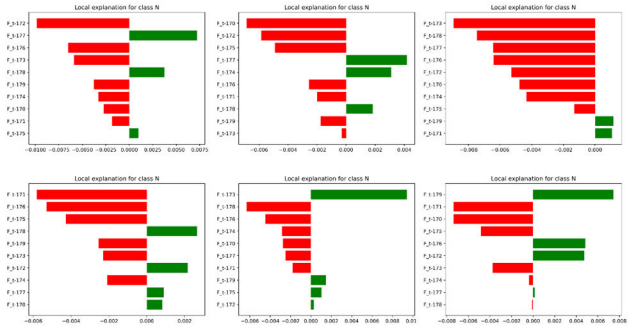

FIGURE 7. LIME explanation outputs across different runs for the same instance. The x axis represents the feature weight while the y axis represents the feature name.

This visual representation emphasizes the instability of LIME explanations, which may be ascribed to the use of locally linear surrogate models with feature selection and the data perturbation approach. For the purpose of fostering confidence and allowing efficient decision-making based on the insights from machine learning models, it is essential to improve the stability and consistency of explanations.

4) INVESTIGATING LOCAL FIDELITY

The lack of local fidelity in the explanations provided while using the Local Interpretable Model-Agnostic Explanation (LIME) approach on Electrocardiography (ECG) signals is one of the major difficulties encountered.

Local fidelity is the degree to which the explanations faithfully represent the behavior of the underlying deep learning model close to the particular instance being described [55]. The lack of local fidelity might make explanations harder to understand and less reliable, which can be problematic in the context of ECG data, where temporal dependencies and distinctive features are essential for classifying arrhythmias.

The data perturbation approach used in LIME includes creating perturbed samples in the vicinity of the original ECG instance to build a local dataset for the interpretable model. Simply adjusting the data points in the case of ECG signals, though, would not be sufficient to fully represent the temporal relationships and distinctive features contained in the signal. ECG signals are time-series data, where each data point corresponds to a specific time interval, and the sequence of data points represents the electrical activity of the heart over time. Perturbing the data independently may disrupt the temporal relationships and lead to explanations that do not accurately reflect the local behavior of the deep learning model for arrhythmia classification.

Furthermore, this disruption in the natural correlation between features can lead to the generation of data points that deviate significantly from the original signal. From the analysis of Fig. 8 it is evident that the LIME technique produces generated data points that differ significantly from the original signal.

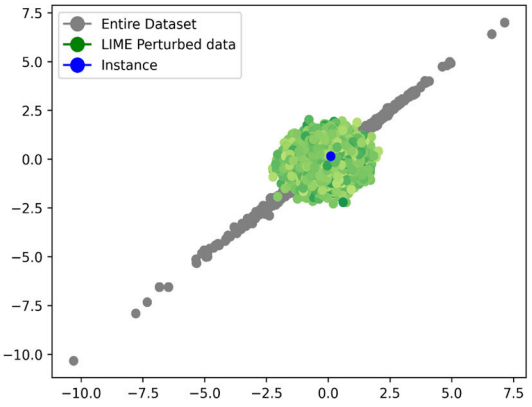

FIGURE 8. The local fidelity of LIME perturbed dataset

These generated data points are represented by the green colour in the figure. The noticeable dissimilarity between the generated data points and the original signal indicates that LIME may not accurately capture the intricate details and nuances of the original signal.

The substantial variance observed between the generated and original data highlights the potential limitation of LIME in capturing intricate patterns within the signal. As such, it becomes evident that overcoming the challenge of local fidelity requires tailored techniques that consider the temporal dynamics and uniqueness of ECG data, ensuring that explanations remain faithful to the underlying model's behavior.

IV. METHODOLOGY OF ENHANCING LIME APPROACH

In this phase, we focus on the methodology of enhancing LIME approach to address specific challenges related to explanation understandability, credibility, stability, scalability, and local fidelity when applied to high-dimensional data, particularly ECG signal data. By identifying these challenges and their underlying reasons, we propose innovative solutions to augment the interpretability and reliability of the explanations provided by LIME.

The primary objective is to generate meaningful explanations that consider the temporal dependencies between features in order to enhance the credibility and interpretability of the results. To tackle these challenges, several modifications have been proposed for the LIME framework. Fig. 9 shows the LIME problems, reasons, and the proposed solutions.

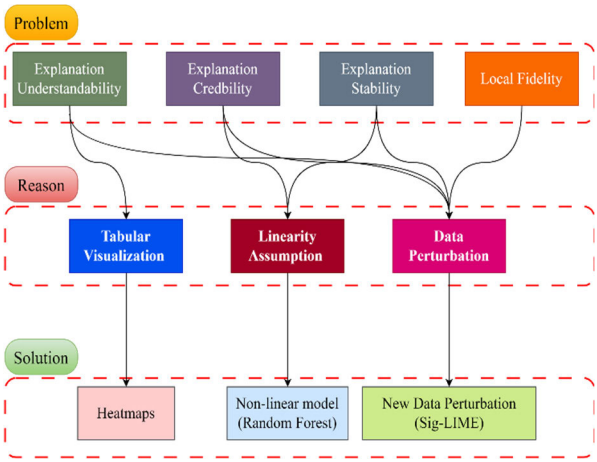

FIGURE 9. Roadmap of LIME problems, reasons, and possible solutions.

A. ENHANCING EXPLANATION UNDERSTANDABILITY THROUGH HEATMAPS

Explanation understandability is a critical aspect of the interpretability of machine learning models, particularly in the context of complex models such as deep learning. Heatmaps provide a visually comprehensible representation of how various characteristics or components contribute to a model's prediction, which makes it easier to understand complicated model behaviors.

Heatmaps are visual depictions that make use of color gradients to show the relative weight or effect of different elements in a particular input instance [56]. In the case of ECG-based cardiac arrhythmia classification, Heatmaps can be used to show the relevance of various time points within an ECG signal in relation to the model's prediction.

By generating heatmaps, LIME enhances its interpretability by presenting a clear and intuitive visual representation of the ECG signal's contribution to the model's decision.

The generation of heatmaps involves assigning color intensities to specific regions of the ECG signal, corresponding to the level of importance of each feature. This can be mathematically represented as follows [57]:

$$Heatmap\,(X) = \sum_{i=1}^{N} w_i.X_i \tag{1}$$

where $Heatmap\,(X)$ represents the heatmap generated for the ECG signal X, N is the total number of features or segments within the ECG signal, w_i denotes the weight assigned to the ith feature, representing its importance, X_i corresponds to the ith feature or segment of the ECG signal.

B. ENHANCING EXPLANATION LINEARITY THROUGH RANDOM FOREST

Explanation credibility is a fundamental requirement for the interpretability of machine learning models, ensuring that the provided explanations accurately reflect the model's behavior. LIME approach, while valuable, can be limited by its inherent linearity assumption. To address this limitation and enhance explanation credibility, the integration of a non-linear model, such as the Random Forest algorithm, presents a promising solution.

LIME generates explanations by fitting a linear model to locally perturbed data points. While this approach is effective in approximating the model's behavior, it may not accurately capture the intricate non-linear relationships present in complex datasets. This linearity assumption can lead to explanations that are overly simplified and do not faithfully represent the true interactions among features [27].

Random Forest is a versatile ensemble learning algorithm known for its ability to capture complex non-linear relationships within data. By integrating Random Forest into the LIME approach, the limitation of linearity assumption can be mitigated. Random Forest can offer more accurate and comprehensive explanations by modeling the intricate feature interactions present in the data, making it a suitable candidate for enhancing explanation credibility [58].

The integration of Random Forest into LIME involves replacing the linear explainer with a Random Forest-based explainer. The Random Forest explainer generates explanations by utilizing the ensemble of decision trees to capture non-linear interactions among features. Mathematically, this can be expressed as follows:

$$Explanation_{RF}\,(X) = \sum_{i=1}^{N} RF_i\,(X_i) \tag{2}$$

where $Explanation_{RF}\,(X)$ denotes the explanation provided by the Random Forest explainer for the input instance X, N represents the number of decision trees in the Random Forest ensemble, $RF_i\,(X_i)$ represents the contribution of the ith decision tree to the explanation for the ith feature or segment X_i of the input.

When applied to the hybrid 1D CNN-GRU model for cardiac arrhythmia classification, the integration of Random Forest with LIME can lead to more faithful explanations. Clinicians and researchers can gain a deeper understanding of the model's decision-making process by observing the non-linear interactions among various segments of the ECG signal. This enhanced explanation credibility contributes to better informed medical decision-making and more robust model validation.

C. ENHANCING DATA PERTURBATION THROUGH SIG-LIME

Data perturbation is a crucial step in LIME approach, as it involves generating local instances for explanation. However, traditional perturbation methods can pose challenges in terms of explanation stability and local fidelity. To address these challenges and enhance the robustness of explanations, novel data perturbation technique (Sig-LIME) is proposed.

Sig-LIME stands as an innovative extension of LIME framework, tailored for signals like ECG signals, where the interpretability of machine learning model predictions is of paramount importance. In the context of signal data, such as ECG, Sig-LIME enriches the data generation and explanation processes of LIME, unraveling deeper insights into model predictions.

At its core, Sig-LIME addresses the limitations of existing methods by striving to generate signal-level explanations that not only pinpoint relevant signal segments but also preserve the temporal intricacies within the data. This becomes especially crucial when understanding models' decisions based on signals, like ECG records, where patterns evolve over time. The overarching goal of Sig-LIME can be summarized in two dimensions: first, identifying the salient signal segments that contribute to predictions, and second, upholding the temporal coherence of the signal during the explanation process.

Signal segmentation forms the foundation of Sig-LIME's approach, partitioning the signal into discrete segments, each open to independent manipulation. The synergy between Gaussian noise and Signal-to-Noise Ratio (SNR) then comes into play, orchestrating the generation of novel heartbeats. This fusion of concepts from the realm of signal processing is instrumental, as Gaussian noise and SNR are time-honored tools in signal analysis, hailing their significance in diverse fields such as audio analysis and communication engineering.

1) GAUSSIAN NOISE

Gaussian noise refers to a type of random variation that is often added to signals to simulate real-world noise. It is characterized by its probability distribution, which follows the Gaussian or normal distribution. This distribution is described by its mean (μ) and standard deviation (σ), where the values of μ and σ determine the characteristics of the noise added to the signal. Gaussian noise effectively

models various sources of uncertainty or interference that can affect signals, such as electronic noise, sensor inaccuracies, and environmental disturbances [59]. Mathematically, the probability density function (PDF) of Gaussian noise is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3}$$

where x represents the noise value, μ is the mean (average) value of the noise, σ is the standard deviation, controlling the spread or dispersion of the noise values.

2) SIGNAL-TO-NOISE RATIO (SNR)

Signal-to-Noise Ratio (SNR) is a quantitative measure used to assess the relative strength of a signal against the presence of noise. It quantifies the clarity of the signal by comparing its magnitude to the magnitude of the noise. SNR is defined as the ratio of the power of the signal and is often expressed in decibels.

The SNR is typically calculated by taking the logarithm (base 10) of the ratio between the power of the signal (P_{signal}) and the power of the noise (P_{noise}), as expressed in the following equation [60].

$$SNR = 10\log_{10}(P_{signal}/P_{noise}) \tag{4}$$

A higher SNR indicates a stronger signal presence relative to noise, signifying better signal quality and increased accuracy in analysis or processing.

In the context of the Sig-LIME technique, Gaussian noise and SNR play a critical role in generating new heartbeats for data perturbation. By introducing controlled Gaussian noise with specific SNR values, Sig-LIME ensures that the perturbed samples maintain the essential features of the original signal while introducing realistic variability. The SNR value governs the trade-off between signal strength and noise level, allowing for the creation of diverse yet meaningful perturbed samples that accurately represent the original signal's local structure.

In the initial phase of the Sig-LIME data perturbation the segmentation of the input signal (S) is conducted. This process involves dividing S into a series of overlapping windows $W = \{w_0, w_1, w_2, \ldots, w_n\}$. Each window w_i is defined as a contiguous subsequence of S, with a fixed length L. The overlap between successive windows is denoted as O.

Mathematically, each window w_i is represented as:

$$w_i = S[i : i + L] \tag{5}$$

where i increments by O in each step, starting from 0. The segmentation concludes upon reaching the end of S is reached, ensuring each segment is of length L.

The subsequent step involves the generation of signal noise. This step is critical in creating synthetic variants of the original signal segments, enabling a robust analysis of the signal's response to various noise conditions. Initially, the function retrieves the signal segments W obtained from S. For each segment w_i, the algorithm computes the quantity of perturbed samples P, which is a function of the total sample size and the number of segments. Mathematically, this is expressed as in formula 6.

$$P = \frac{L}{len(W)} \tag{6}$$

The core of the noise generation phase is designed to introduce Gaussian noise into each segment based on a defined Signal-to-Noise Ratio (SNR). For a given SNR value, the corresponding noise power P_n is computed using the following.

$$P_n = \frac{signal\ Power}{10^{(SNR/10)}} \tag{7}$$

where Signal Power is the sum of the squared values of the signal segment. This formulation ensures that the noise introduced is proportional to the signal's power, adhering to the desired SNR level. The resulting noise-infused segments are compiled into an array Ps.

The next critical step is the Data Generation process. This phase focuses on synthesizing a comprehensive dataset from the noise-augmented segments Ps, facilitating the evaluation and explanation of the signal processing model's behaviour. Firstly, we create an array of zeros ($NewData$) with dimensions of the number of sample size (Z) and the Signal length $len(S)$.

The $NewData$ array functions as a foundational matrix where the noise-enhanced segments from Ps are systematically incorporated. This integration is performed through an iterative process, wherein each individual segment from Ps is methodically embedded into the $NewData$ array. The precise allocation of these segments within $NewData$ is controlled by specific row indices r_{start} and r_{end}, and column indices c_{start} and c_{end}. The mathematical formulation of this embedding process is expressed as:

$$NewData[r_{start} : r_{end}, c_{start} : c_{end}] = Ps[i] \tag{8}$$

Fig. 10 presents an overarching view of the Sig-LIME methodology, underscoring its commitment to producing explanations that not only maintain credibility but also respect the temporal dependencies ingrained within signal data.

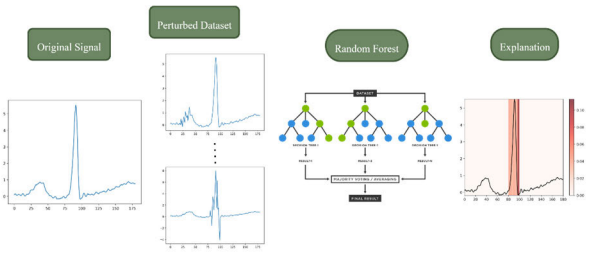

FIGURE 10. Sig-LIME procedure to generate explanations.

Sig-LIME's ingenuity shines through in its intricate procedure, harmonizing signal segmentation and noise manipulation to generate meaningful explanations. This

Algorithm 1 Pseudocode for Sig-LIME Algorithm

```
Input: DLModel (model), Signal (S),
SampleSize (Z), SengmetLenght (L),
Overlap(O)
Output: Explanation
Start Algorithm (Sig-LIME)
 Phase 1: Split ECG Signal into segments
Function SegmentSignal(S, L, O):
    Initialize W as empty list for
Segments
   FOR start FROM 0 TO (length(S) - L)
 STEP O:
        end = Min(start + L, length(S))
        IF end = length(S) THEN
          EXIT LOOP
          w_i = S[start:end]
          APPEND w_i to W
    RETURN W
End;//Function
   Phase 2, Generate noise based on each
                    segment
Function GenerateNoise(S, L, SNR):
    Initialize Ps as empty list for
PerturbedSegments
    W = SegmentSignal(S, L, O)
    P = L / length(W)
    FOR each i in W:
        FOR each j in SNR:
            noise =
GenerateNoiseFor Segment (i, j, (P/10,
length(i)))
            Append noise to Ps
    Return Ps
End;// Function
    Phase 3, Generate Perturbed Dataset
Function DataGenerator(Z, S):
    Ps = GenerateNoise(S, L)
    NF = Number of features in S
    NData = Zero Matrix of size (Z, NF)
    RStart, REnd, CStart, CEnd = 0
    FOR each perturbed segment i in Ps:
        REnd = RStart + length(Ps[i])
        CEnd = CStart + Number of
features in Ps[i]
        NewData[RStart:REnd, CStart:CEnd]
+= Ps[i]
        Update RStart, CStart, REnd, CEnd
        Reset parameters if needed
     NData += S
     Return NData
End;// Function
     Phase 4, Generate explanations
Initialize NewData, Labels, Neighborhood,
Importance
NewData = DataGenerator(Z, S)
Labels = model.predict(NewData)
Neighborhood = Calculate pairwise
distances between S and NewData
RFModel = RandomForestClassifier()
RFModel.fit(NewData, Labels,
Neighborhood)
Importance = RFModel.feature_importances
PlotHeatmap(Importance, Signal S)
End;//Algorithm
```

technique ensures that only specific segments of the signal undergo transformation, leaving the rest untouched, thus upholding the temporal relationships inherent in the signal.

The proposed Sig-LIME algorithm is outlined in Algorithm 1, presented as pseudocode.

V. RESULTS

The Sig-LIME approach is an improvement to the data generation and explanation methods used by LIME that is designed exclusively for producing signal-level explanations for models that accept signals as input, such as ECG signals.

Signal segmentation and noise production are the two crucial elements of the data generating method used by Sig-LIME. By breaking the ECG signal up into smaller chunks, signal segmentation enables independent manipulation of each segment.

Following signal segmentation, Sig-LIME uses the Signal Noise Ratio (SNR) and Gaussian random noise generating algorithms. These methods insert controlled noise into each segment of the signal. The other feature values in the segment, which preserve the temporal connections between features, are significant. This strategy helps the extraction of important insights from challenging ECG datasets, which leads to enhanced decision-making processes. Fig. 11 shows examples of the data generated by Sig-LIME.

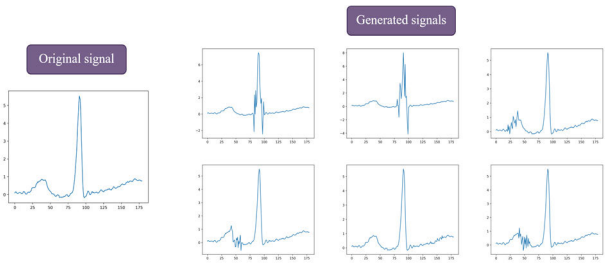

FIGURE 11. Samples of the data generated by Sig-LIME

As a result, Sig-LIME achieves a substantial enhancement in the neighborhood of the original data, contributing to more accurate and credible explanations. The visual representation in Fig. 12 vividly depicts the synthetic data generated by Sig-LIME, offering a transparent illustration of the improved data generation process in action. This enhancement underscores the efficacy of Sig-LIME in preserving the key features of the original signal while introducing controlled noise, thereby augmenting the quality of explanations provided by the model.

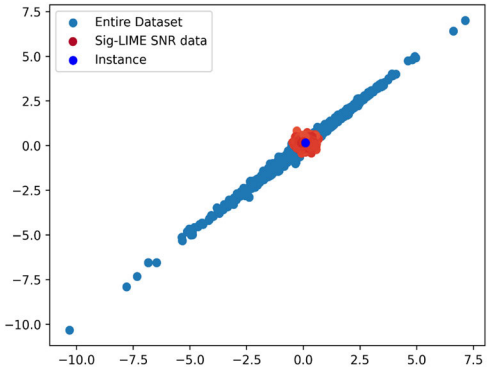

FIGURE 12. Synthetic dataset generated by Sig-LIME

Heatmaps and a random forest model are used in Sig-LIME. Together, these elements produce insightful justifications for the predictions produced by models that use ECG data as input. Fig 13 illustrates the results of the Sig-LIME explanation approach by using heatmaps to present the importance of features in a visual manner. It is clear that the heatmaps emphasize the characteristics closest to the QRS complex the most. This finding validates the argument that the QRS complex is essential to the model's predictions and establishes the validity of the Sig-LIME technique for gathering significant signal-level data. These results demonstrate the efficiency of Sig-LIME in producing explanations that stress the QRS complex and offer insightful information about the model's decision-making process.

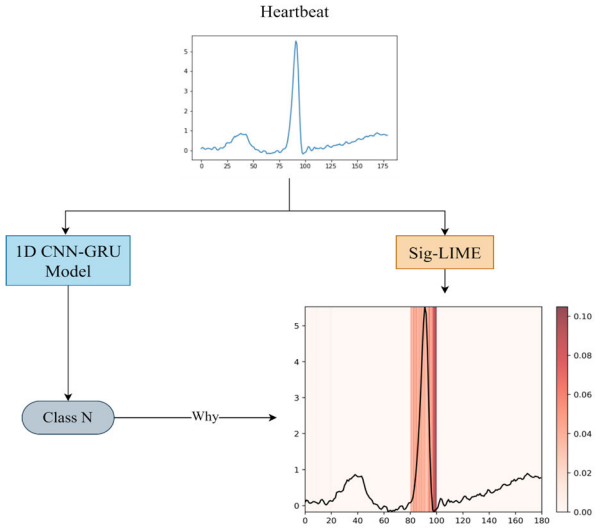

FIGURE 13. Sig-LIME explanation outcome.

Fig 14 presents a wide range of samples that the inquiry examined and the significant findings that came from in-depth research and analysis.

The use of heatmaps improves the explanation's readability and visual representation, allowing for a deeper comprehension of the variables impacting the model's predictions.

A. COMPARISON WITH BASELINE LIME

As previously discussed, LIME faces challenges in terms of explanation understandability, credibility, instability, and local fidelity due to its random feature perturbation data generation technique and the linear assumption. To address these limitations, the proposed method, namely Sig-LIME, introduced more realistic and informative explanations.

1) COMPARISON IN TERMS OF UNDERSTANDABILITY

One of the fundamental objectives of enhancing the interpretability of explanations is to ensure their understandability across diverse audiences. In this context, we delve into a comprehensive comparison between the proposed method and the baseline LIME approach in terms of the understandability of the generated explanations.

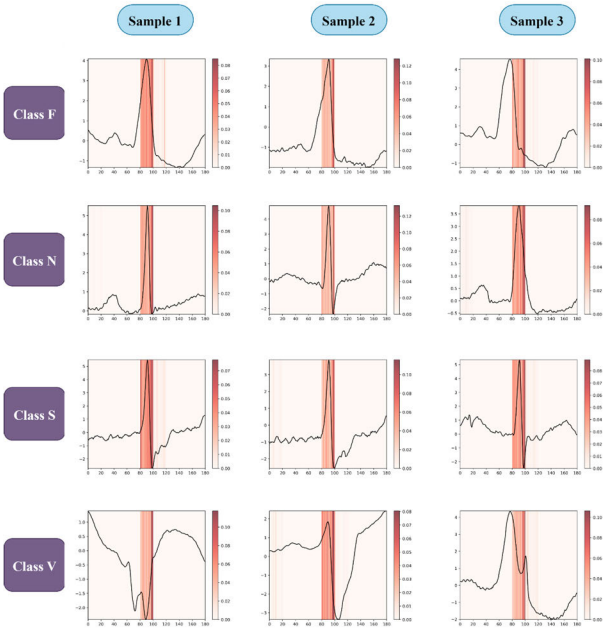

FIGURE 14. Diverse arrhythmia types explored through Sig-LIME explanation.

The inclusion of the heatmap visualization technique stands out as a pivotal factor contributing to the improved understandability of explanations. This technique facilitates the overlay of feature weights onto the ECG signal, providing a more intuitive representation of their significance. The intensity of the red color reflects the relative importance of features, where darker hues signify greater importance. This visual representation offers a compelling comparison of explanation outcomes among LIME and Sig-LIME.

While LIME employs a conventional approach that combines textual and visual elements for conveying explanations, its effectiveness diminishes when dealing with intricate signal data. To overcome this challenge, the integration of heatmap techniques within Sig-LIME is paramount. Heatmaps emerge as powerful tools for visually representing feature contributions within signals, which in turn augments the comprehensibility of the explanations. By emphasizing critical areas, like the QRS complex, heatmaps provide insights into influential features, enhancing the interpretation process.

Fig. 15 effectively illustrates the contrasting explanation outcomes of LIME in comparison to the proposed Sig-LIME method.

The latter consistently highlight regions proximate to the QRS complex, aligning with the recognized significance of this region in arrhythmia analysis. In contrast, LIME's tendency to emphasize random areas without a specific focus becomes evident. Furthermore, the heatmap sidebar within the same figure signifies the heightened confidence exhibited by Sig-LIME in their chosen features when compared to LIME. This confidence is visibly pronounced through significantly higher importance scores assigned to specific features in Sig-LIME in comparison to LIME.

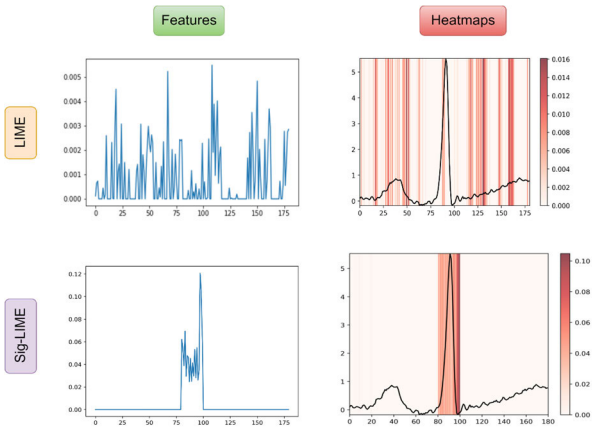

FIGURE 15. Comparison of the explanation outcomes of LIME with the proposed Sig-LIME.

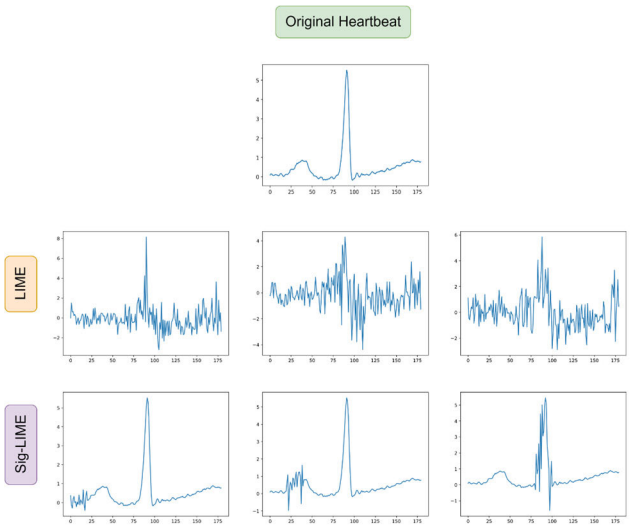

FIGURE 16. Comparison of data Generation between LIME, and Sig-LIME.

Through the integration of heatmaps, the proposed methods distinctly elevate the interpretability and clarity of the explanations. This enhanced understandability empowers researchers, practitioners, and domain experts to gain a deeper understanding of the intricate relationships within ECG signals, facilitating informed decision-making and contributing to the advancement of cardiac arrhythmia classification practices.

2) COMPARISON IN TERMS OF CREDIBILITY

The evaluation of credibility emerges as a pivotal dimension when contrasting the proposed Sig-LIME method with the established LIME technique. LIME's lack of credibility, primarily stemming from its perturbation technique and reliance on linearity assumptions, triggers a comprehensive assessment of these methods' ability to generate explanations that inspire trust and reliance.

The veracity of LIME's explanations is notably undermined by its random feature perturbation technique and adherence to linear assumptions. This inherent limitation has spurred skepticism regarding the credibility of LIME's explanations.

To substantiate these concerns, Fig. 16 unfolds as a visual testament, depicting samples that underscore the lack of credibility stemming from LIME's perturbation technique. This figure also affords a comparative analysis of data generated by LIME, and Sig-LIME, further accentuating the credibility concerns surrounding LIME.

Sig-LIME undertake ingenious strategies to tackle the credibility challenge, harnessing the inherent nature of ECG signals to produce realistic explanations. Sig-LIME partitions heartbeats into segments and leverages Gaussian noise and Signal-to-Noise Ratio (SNR) to generate data points that capture the underlying signal characteristics.

The credibility quandary posed by LIME's linear model is further examined, considering the high dimensionality and intricacy of ECG signals. In response, the incorporation of Random Forest (RF) models within Sig-LIME emerges as a potent solution. RF's aptitude for handling both linear and non-linear relationships, without the need for excessive hyperparameter tuning, imparts enhanced credibility to the explanations provided by these methods.

Intriguingly, Fig. 17 materializes as a critical point of comparison, contrasting the feature weights derived from LIME with those generated by Sig-LIME. This visual representation encapsulates LIME's tendency to choose random areas with wavering confidence, juxtaposed with the proposed methods' consistent focus on the QRS complex region. This alignment with the known importance of the QRS complex in arrhythmia analysis reiterates the credibility of the explanations offered by Sig-LIME.

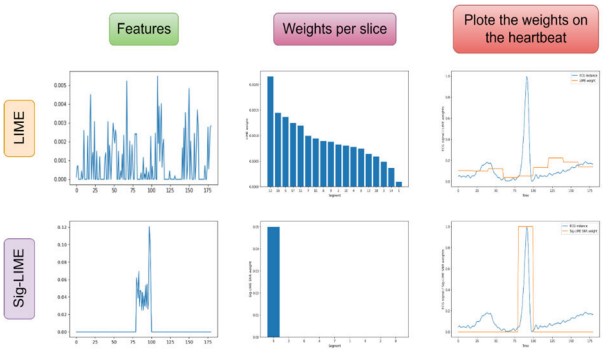

FIGURE 17. Comparison of the feature weights generated by LIME and the proposed Sig-LIME.

3) COMPARISON IN TERMS OF STABILITY

The examination of stability constitutes another vital dimension in our comparative analysis of LIME, and Sig-LIME. Stability pertains to the degree of consistency exhibited by the explanations across various perturbations or multiple runs of the explanation process. This assessment is visually

represented in Fig. 18, offering insights into the stability characteristics of the respective methods.

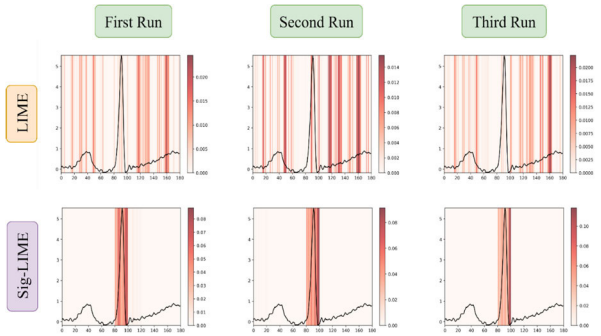

FIGURE 18. Comparison of stability between LIME explanation outcomes and Sig-LIME.

Turning our attention to LIME, it becomes evident that the generated explanations lack stability. The model tends to highlight disparate regions within the cardiac signal upon each execution or data alteration. This variability arises due to LIME's perturbation approach, introducing inconsistencies in the explanations.

In contrast, Sig-LIME manifest superior stability. The method consistently emphasizes the vicinity of the QRS complex in its explanations, regardless of the number of executions or data manipulations. This stability is a result of the distinctive strategies employed by Sig-LIME.

To quantitatively evaluate the stability of the explanations, we employed an analysis of variance (ANOVA) framework. Our experimental design involved running LIME, and Sig-LIME multiple times (specifically, three runs) and recording the feature weights obtained in each run. Subsequently, ANOVA was applied to ascertain the degree of similarity between these runs for each method.

The results yielded unequivocal evidence of a marked improvement in stability attributed to our proposed methodologies. In the case of LIME, an f-statistic of 0.76 was observed, accompanied by a corresponding p-value of 0.4689.

On the other hand, Sig-LIME displayed an f-statistic of 0.0, indicating a complete absence of variation between runs. Moreover, Sig-LIME yielded p-values of 1, which is the highest possible p-value and indicates a strong statistical agreement between the runs.

This outcome demonstrates an exceptionally high level of stability in the explanations produced by Sig-LIME. In other words, the feature weights in these techniques remained virtually identical across all iterations, providing strong evidence of their stability.

LIME's instability is predominantly attributed to its uniform perturbation technique involving Gaussian noise addition. This simplistic perturbation method may result in explanations that disrupt the inherent data structure and coherence.

In contrast, Sig-LIME mitigates this instability by segmenting the signal and introducing Gaussian noise based on the Signal-to-Noise Ratio (SNR). This approach preserves the temporal interdependencies within the signal, thereby yielding explanations that are notably more reliable and consistent.

Furthermore, the assumption of linear relationships between features contributes to LIME's instability. This assumption often proves untenable for complex real-world data, resulting in inconsistent feature selection and, consequently, explanations.

4) COMPARISON IN TERMS OF LOCAL FIDELITY

The concept of local fidelity takes center stage as we delve into a comparative analysis of LIME and Sig-LIME, focusing on their capacity to faithfully represent the behavior of the model near a specific instance. This analysis becomes especially pertinent in the context of ECG signals, where temporal dependencies and unique characteristics are paramount for accurate arrhythmia classification. The presence of local fidelity serves as a cornerstone for the generation of reliable explanations.

Fig. 19 compares the local fidelity of Sig-LIME and LIME on a synthetic dataset. Local fidelity is measured by the average Euclidean distance between the predictions of the LIME model and the original model. As the figure shows, Sig-LIME has a lower average Euclidean distance than LIME with 0.49 for Sig-LIME compared to 17.24 for LIME. This means that Sig-LIME is better at generating data that is more similar to the original data, on average. Sig-LIME shines as it makes substantial strides in enhancing the local fidelity of the generated explanations.

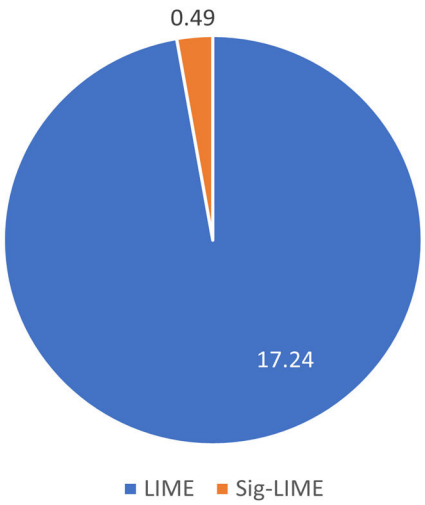

FIGURE 19. Average Euclidean Distance of the data generated by LIME and Sig-LIME to the original data point.

In this pursuit, Sig-LIME shine as they make substantial strides in enhancing the local fidelity of the generated explanations. This remarkable progress is elegantly portrayed in Fig. 20, where each data point's color signifies its

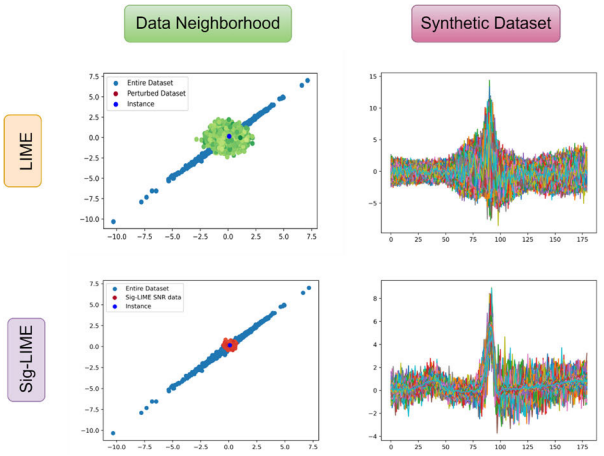

FIGURE 20. Comparison of the local fidelity between LIME, and Sig-LIME

proximity to the original ECG heartbeat data point (depicted as a blue dot). The spectrum of colors, ranging from light to dark, symbolizes the degree of dissimilarity between the generated data points and the original instance.

Notably, within LIME's framework, the green-colored data points reveal a noticeable deviation from the original instance. This departure underscores a limitation in local fidelity, where LIME's generated explanations might not entirely capture the essence of the model's decision-making process, raising concerns about their reliability.

In stark contrast, Sig-LIME emerges as a beacon of high local fidelity. The data points it generates closely echo the characteristics of the original data point. This feat is achieved through Sig-LIME's innovative approach of integrating Gaussian noise and Signal-to-Noise Ratio (SNR) into data point generation. By skillfully injecting controlled noise and signal variance, Sig-LIME ensures that the generated data points encapsulate the intrinsic temporal dependencies found within ECG signals.

The enhancement in local fidelity offered by Sig-LIME underscores their efficacy in delivering explanations that authentically mirror the model's behavior within the intricate landscape of ECG signals. This advancement doesn't merely bolster the trustworthiness of the explanations; it also aligns them with the nuanced patterns that define cardiac arrhythmia classification.

VI. DISCUSSION

The development of Sig-LIME marks a significant advancement in the interpretability of machine learning models, especially in the context of signal data such as ECG signals. This enhanced version of the Local Interpretable Model-agnostic Explanations (LIME) algorithm effectively addresses the intricate challenges inherent in interpreting complex models. Its application is especially crucial in the medical domain, where the interpretation of ECG recordings and similar signal data is paramount.

A key strength of Sig-LIME lies in its ability to significantly improve the interpretability of machine learning models. This advancement is vital in contexts where the accuracy and trustworthiness of model predictions are critical. Sig-LIME's capacity to offer clearer, more intuitive explanations of model decisions facilitates a deeper, more comprehensive understanding of complex models. This feature is invaluable in high-stakes applications like medical diagnostics, where precise and reliable interpretations can have profound implications.

The algorithm's unique approach to maintaining temporal relationships in signal data is particularly beneficial for time-series analysis. This preservation ensures not only the accuracy of the interpretations but also their contextual relevance. Such an approach is essential for maintaining the integrity of time-sensitive data analyses.

In terms of stability, Sig-LIME represents a substantial enhancement over the conventional LIME model. The increased stability in its explanations contributes significantly to the algorithm's credibility and the consistency of its interpretative outcomes. Additionally, Sig-LIME advances the local fidelity of explanations, effectively tackling a common challenge in machine learning interpretability. The integration of novel data generation techniques and the application of heatmaps and random forest models in Sig-LIME enable a more detailed and accessible understanding of model behaviors in signal data processing. These attributes, when juxtaposed with the baseline LIME algorithm, underscore Sig-LIME's superiority in providing nuanced and comprehensive interpretations.

While Sig-LIME introduces a groundbreaking methodology for augmenting interpretability in signal-based machine learning models, it is imperative to acknowledge certain inherent limitations. One notable demerit lies in the sensitivity of Sig-LIME to the choice of parameters, particularly in the signal perturbation phase. The effectiveness of Sig-LIME can be influenced by the choice of parameters such as signal-to-noise ratio (SNR) and segment size. Challenges in accurately determining these parameters may expose the algorithm to suboptimal configurations, potentially impacting the fidelity of explanations produced.

While this study marks a significant stride forward, addressing the algorithm's sensitivity becomes a crucial focus for future research. Subsequent investigations could involve meticulous evaluations of Sig-LIME across diverse signal data types beyond ECG signals. Additionally, future work will focus on enhancing the algorithm's robustness by exploring automated parameter tuning mechanisms such as HSCATSO [61] and incorporating adaptive strategies to mitigate its sensitivity across varying signal characteristics.

VII. CONCLUSION

This study has tackled the challenges of interpreting signal data, specifically ECG signals, using the LIME technique. We systematically investigated the limitations of LIME, including its perturbation technique and linear assumptions, and introduced Sig-LIME as a solution. Through a thorough comparison, we highlighted Sig-LIME's significant

improvements over LIME. Our analysis revealed that Sig-LIME, with its innovative use of Gaussian noise and Signal-to-Noise Ratio (SNR), addressed credibility, stability, and local fidelity issues. The integration of heatmap visualization enhanced the understandability of explanations by overlaying feature weights on ECG signals, emphasizing critical regions like the QRS complex. Furthermore, by adopting Random Forest instead of ridge regression, Sig-LIME captured non-linear relationships in complex ECG data, elevating the credibility of explanations. While marking a significant advancement, this study recognizes avenues for future research. Further refinements in the Sig-LIME methodology could involve evaluating its performance across various signal data types beyond ECG signals. Additionally, the study acknowledges the limitation of Sig-LIME's sensitivity to parameter choices, and future work will focus on addressing this challenge through automated parameter tuning mechanisms and adaptive strategies, ensuring robust performance across diverse signal characteristics. Furthermore, exploring the integration of Sig-LIME into existing clinical workflows and assessing its impact on decision-making processes will provide valuable insights into its real-world applicability.

ACKNOWLEDGMENT

The authors would like to thank the technical support provided by the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

REFERENCES

[1] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *Proc. Workshop Explainable AI(IJCAI)*, 2017, pp. 8–13. [Online]. Available: http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf

[2] A. Pfob, B. J. Mehrara, J. A. Nelson, E. G. Wilkins, A. L. Pusic, and C. Sidey-Gibbons, "Towards patient-centered decision-making in breast cancer surgery: Machine learning to predict individual patient-reported outcomes at 1-year follow-up," *Ann. Surgery*, vol. 277, no. 1, pp. e144–e152, 2023.

[3] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, no. 6, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.

[4] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: Definitions, methods, and applications," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 44, pp. 22071–22080, 2019, doi: 10.1073/pnas.1900654116.

[5] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "Interpretability in healthcare: A comparative study of local machine learning interpretability techniques," *Comput. Intell.*, vol. 37, no. 4, pp. 1633–1650, Nov. 2021.

[6] T. Abdullah, W. Ali, S. Malebary, and A. A. Ahmed, "A review of cyber security challenges attacks and solutions for Internet of Things based smart home," *Int. J. Comput. Sci. Netw. Secur.*, vol. 19, no. 9, p. 139, 2019.

[7] T. A. Abdullah, W. Ali, and R. Abdulghafor, "Empirical study on intelligent Android malware detection based on supervised machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, 2020.

[8] A. J. London, "Artificial intelligence and black-box medical decisions: Accuracy versus explainability," *Hastings Center Rep.*, vol. 49, no. 1, pp. 15–21, 2019.

[9] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 10, no. 5, p. e1379, 2020.

[10] T. A. A. Abdullah, M. S. M. Zahid, and W. Ali, "A review of interpretable ML in healthcare: Taxonomy, applications, challenges, and future directions," *Symmetry*, vol. 13, no. 12, p. 2439, Dec. 2021.

[11] V. Haunschmid, E. Manilow, and G. Widmer, "audioLIME: Listenable explanations using source separation," *Expert Rev. Cardiovascular Therapy*, vol. 18, no. 2, pp. 77–84, 2020, doi: 10.1080/14779072.2020.1732208.

[12] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez, "Explainable artificial intelligence (XAI) on timeseries data: A survey," 2021, *arXiv:2104.00950*.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[14] A. Sangroya, M. Rastogi, C. Anantaram, and L. Vig, "Guided-LIME: Structured sampling based hybrid approach towards explaining blackbox machine learning models," in *Proc. CIKM (Workshops)*, 2020.

[15] M. I. Al-Hiyali, A. J. Ishak, H. Harun, S. A. Ahmad, and W. A. W. Sulaiman, "A review in modification food-intake behavior by brain stimulation: Excess weight cases," *NeuroQuantology*, vol. 16, no. 12, pp. 86–97, 2018.

[16] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, and H. Gamboa, "Interpretable heartbeat classification using local model-agnostic explanations on ECGs," *Comput. Biol. Med.*, vol. 133, Jun. 2021, Art. no. 104393, doi: 10.1016/j.compbiomed.2021.104393.

[17] X. Li, H. Xiong, X. Li, X. Zhang, J. Liu, H. Jiang, Z. Chen, and D. Dou, "G-LIME: Statistical learning for local interpretations of deep neural networks using global priors," *Artif. Intell.*, vol. 314, Jan. 2023, Art. no. 103823.

[18] Z. Zhou, G. Hooker, and F. Wang, "S-LIME: Stabilized-LIME for model explanation," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2021, pp. 2429–2438.

[19] G. Visani, E. Bagli, and F. Chesani, "OptiLIME: Optimized LIME explanations for diagnostic computer algorithms," in *Proc. CEUR Workshop*, vol. 2699, 2020.

[20] S. M. Shankaranarayana and D. Runje, "ALIME: Autoencoder based approach for local interpretability," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Lecture Notes in Computer Science), vol. 11871, 2019, pp. 454–463, doi: 10.1007/978-3-030-33607-3_49.

[21] E. Kiymaç and Y. Kaya, "A novel automated CNN arrhythmia classifier with memory-enhanced artificial hummingbird algorithm," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 119162.

[22] A. Vadillo-Valderrama, R. Goya-Esteban, R. P. Caulier-Cisterna, A. García-Alberola, and J. L. Rojo-Álvarez, "Differential beat accuracy for ECG family classification using machine learning," *IEEE Access*, vol. 10, pp. 129362–129381, 2022.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 1527–1535, 2018.

[24] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[25] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[26] R. Gaudel, L. Galárraga, J. Delaunay, L. Rozé, and V. Bhargava, "s-LIME: Reconciling locality and fidelity in linear explanations," in *Proc. Int. Symp. Intell. Data Anal.* Springer, 2022, pp. 102–114.

[27] G. Visani, E. Bagli, and F. Chesani, "OptiLIME: Optimized LIME explanations for diagnostic computer algorithms," *Heart*, vol. 106, no. 5, pp. 318–320, 2020, doi: 10.1136/heartjnl-2019-316030.

[28] T. Botari, F. Hvilshøj, R. Izbicki, and A. C. P. L. F. de Carvalho, "MeLIME: Meaningful local explanation for machine learning models," 2020, *arXiv:2009.05818*.

[29] P. Hall, N. Gill, M. Kurka, W. Phan, and A. Bartz, "Machine learning interpretability with H_2O driverless AI," in *Machine Learning Interpretability With H_2O Driverless AI*, 1st ed., 2019, pp. 1–40. [Online]. Available: http://docs.h2o.ai

[30] L. Hu, J. Chen, V. N. Nair, and A. Sudjianto, "Locally interpretable models and effects based on supervised partitioning (LIME-SUP)," *J. Amer. Heart Assoc.*, vol. 9, no. 4, 2018, doi: 10.1161/jaha.119.013924.

[31] I. Ahern, A. Noack, L. Guzman-Nateras, D. Dou, B. Li, and J. Huan, "NormLime: A new feature importance metric for explaining deep neural networks," 2019, *arXiv:1909.04200*.

[32] M. Rehman Zafar and N. Mefraz Khan, "DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," 2019, *arXiv:1906.10263*.

[33] J. Rabold, M. Siebers, and U. Schmid, "Explaining black-box classifiers with ILP–empowering LIME with Aleph to approximate non-linear decisions with relational rules," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (LNAI), vol. 11105, 2018, pp. 105–117, doi: 10.1007/978-3-319-99960-9_7.

[34] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney, "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review," *Appl. Sci.*, vol. 11, no. 11, p. 5088, May 2021.

[35] T. A. A. Abdullah, M. S. M. Zahid, W. Ali, and S. U. Hassan, "B-LIME: An improvement of LIME for interpretable deep learning classification of cardiac arrhythmia from ECG signals," *Processes*, vol. 11, no. 2, p. 595, Feb. 2023. [Online]. Available: https://www.mdpi.com/2227-9717/11/2/595

[36] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," 2016, *arXiv:1602.02410*.

[37] R. S. Andersen, A. Peimankar, and S. Puthusserypady, "A deep learning approach for real-time detection of atrial fibrillation," *Expert Syst. Appl.*, vol. 115, pp. 465–473, Jan. 2019, doi: 10.1016/j.eswa.2018.08.011.

[38] S. U. Hassan, M. S. M. Zahid, T. A. A. Abdullah, and K. Husain, "Classification of cardiac arrhythmia using a convolutional neural network and bi-directional long short-term memory," *Digit. Health*, vol. 8, pp. 1–13, 2022, doi: 10.1177/20552076221102766.

[39] A. A. Ahmed, W. Ali, T. A. A. Abdullah, and S. J. Malebary, "Classifying cardiac arrhythmia from ECG signal using 1D CNN deep learning model," *Mathematics*, vol. 11, no. 3, p. 562, Jan. 2023.

[40] M. I. Al-Hiyali, N. Yahya, I. Faye, and Z. Khan, "Autism spectrum disorder detection based on wavelet transform of BOLD fMRI signals using pre-trained convolution neural network," *Int. J. Integr. Eng.*, vol. 13, no. 5, pp. 49–56, Jul. 2021.

[41] A. Fred Agarap, "Deep learning using rectified linear units (ReLU)," 2018, *arXiv:1803.08375*.

[42] B. Jason, "A gentle introduction to batch normalization for deep neural networks—MachineLearningMastery," in *Machine Learning Mastery*, 2019.

[43] C. Garbin, X. Zhu, and O. Marques, "Dropout vs. Batch normalization: An empirical study of their impact to deep learning," *Multimedia Tools Appl.*, vol. 79, nos. 19–20, pp. 12777–12815, May 2020.

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Sep. 2014.

[45] M. A. Deif, A. A. Solyman, M. A. Kamarposhti, S. S. Band, and R. E. Hammam, "A deep bidirectional recurrent neural network for identification of SARS-CoV-2 from viral genome sequences," *Math. Biosci. Eng.*, vol. 18, no. 6, pp. 8933–8950, 2021.

[46] M. Zulqarnain, R. Ghazali, M. G. Ghouse, and M. F. Mushtaq, "Efficient processing of GRU based on word embedding for text classification," *JOIV: Int. J. Inform. Vis.*, vol. 3, no. 4, pp. 377–383, 2019.

[47] M. I. Al-Hiyali, N. Yahya, I. Faye, M. S. Al-Quraishi, and A. Al-Ezzi, "Principal subspace of dynamic functional connectivity for diagnosis of autism spectrum disorder," *Appl. Sci.*, vol. 12, no. 18, p. 9339, Sep. 2022.

[48] Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, "Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network," *Inf. Fusion*, vol. 53, pp. 174–182, Jan. 2020.

[49] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7.

[50] T. A. A. Abdullah, M. S. B. M. Zahid, T. B. Tang, W. Ali, and M. Nasser, "Explainable deep learning model for cardiac arrhythmia classification," in *Proc. Int. Conf. Future Trends Smart Communities (ICFTSC)*, Dec. 2022, pp. 87–92.

[51] M. Zdravković, I. Ćirić, and M. Ignjatović, "Explainable heat demand forecasting for the novel control strategies of district heating systems," *Annu. Rev. Control*, vol. 53, pp. 405–413, Jan. 2022.

[52] A. Ferraro, A. Galli, V. Moscato, and G. Sperlì, "Evaluating eXplainable artificial intelligence tools for hard disk drive predictive maintenance," *Artif. Intell. Rev.*, vol. 56, no. 7, pp. 7279–7314, 2023.

[53] M. I. Al-Hiyali, N. Yahya, I. Faye, A. Sadiq, and M. N. M. Saad, "Detection of Alzheimer's disease using dynamic functional connectivity patterns in resting-state fMRI," in *Proc. Int. Conf. Future Trends Smart Communities (ICFTSC)*, Dec. 2022, pp. 49–54.

[54] T. Botari, F. Hvilshøj, R. Izbicki, and A. C. de Carvalho, "MeLIME: Meaningful local explanation for machine learning models," *Open Heart*, vol. 7, no. 1, 2020, doi: 10.1136/openhrt-2020-001297.

[55] M. U. Ansari, "ULIME: Uniformly weighted local interpretable model-agnostic explanations for image classifiers," Tech. Rep., 2022.

[56] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[57] S. Busygin, O. A. Prokopyev, and P. M. Pardalos, "Feature selection for consistent biclustering via fractional 0–1 programming," *J. Combinat. Optim.*, vol. 10, pp. 7–21, Aug. 2005.

[58] N. M. Abdulkareem and A. M. Abdulazeez, "Machine learning classification based on radom forest algorithm: A review," *Int. J. Sci. Bus.*, vol. 5, no. 2, pp. 128–142, 2021.

[59] Z.-Q. Chen and Y. Hu, "Solvability of parabolic Anderson equation with fractional Gaussian noise," *Commun. Math. Statist.*, vol. 11, no. 3, pp. 563–582, Sep. 2023.

[60] W. J. Lira de Queiroz, D. Brito Teixeira de Almeida, F. Madeiro, and W. T. A. Lopes, "Signal-to-noise ratio estimation for M-QAM signals in $\eta - \mu$ and $\eta - \mu$ fading channels," *EURASIP J. Adv. Signal Process.*, vol. 2019, p. 20, Mar. 2019.

[61] H. Shutari, T. Ibrahim, N. B. M. Nor, H. Q. A. Abdulrab, N. Saad, and Q. Al-Tashi, "Coordination of enhanced control schemes for optimal operation and ancillary services of grid-tied VSWT system," *IEEE Access*, vol. 11, pp. 43520–43535, 2023.

TALAL ALI AHMED ABDULLAH received the B.Sc. degree in computer science from the Faculty of Science, Taiz University, Yemen, in 2014, and the M.Sc. degree in information technology from the Kulliyyah of Information and Communication Technology, International Islamic University Malaysia (IIUM), Malaysia, in 2019. He is currently pursuing the Ph.D. degree with the Computer and Information Sciences Department, Universiti Teknologi PETRONAS (UTP), Malaysia. His research interest includes machine learning techniques and their applications.

MOHD SOPERI MOHD ZAHID (Member, IEEE) received the Ph.D. degree in computer science from the University of Wisconsin–Milwaukee, in 2009. He is currently an Associate Professor with the Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Malaysia. Prior to that, he was a Faculty Member of the Faculty of Computing, Universiti Teknologi Malaysia. He has published papers in numerous journals and conference proceedings. His main research interests include computer network failure recovery and machine learning for software defined network security and healthcare.

AHMAD F. TURKI is currently an Assistant Professor in biomedical engineering with the Department of Electrical Engineering and Computer Engineering, Faculty of Engineering, King Abdulaziz University (KAU), Jeddah, Saudi Arabia.

WALEED ALI received the B.Sc. degree in computer science from the Faculty of Science, Taiz University, Yemen, in 2005, and the M.Sc. and Ph.D. degrees in computer science from the Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia, in 2009 and 2012, respectively. He has been an Associate Professor with the Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, since September 2013. His research interests include intelligent web systems, cybersecurity, and soft computing and machine learning.

AHMAD A. JIMAN received the B.Sc. degree in electrical engineering from King Abdulaziz University, in 2012, the M.Sc. degree in biomedical engineering, in 2015, and the Ph.D. degree from the University of Michigan, Ann Arbor, MI, USA, in 2020. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, King Abdulaziz University, Jeddah, Saudi Arabia. His dissertation was on interfacing with autonomic nerves to regulate blood glucose levels. His current research interests include biosensors, neural engineering, neuroscience, and diabetes.

MOHAMMED J. ABDULAAL is currently an Associate Professor with the Department of Electrical and Computer Engineering, King Abdulaziz University, Saudi Arabia. His research interests include signal processing and machine learning of biomedical systems, Hajj research, traffic control systems, cybersecurity, and various image processing applications.

NEBRAS M. SOBAHI is currently an Associate Professor with the Department of Electrical and Computer Engineering, King Abdulaziz University, Saudi Arabia. His research interests include nano/microfabrication, MEMS, microfluidics, BioMEMS, and signal and image processing.

EYAD T. ATTAR is currently an Assistant Professor in biomedical engineering with the Department of Electrical Engineering and Computer Engineering, Faculty of Engineering, King Abdulaziz University (KAU), Jeddah, Saudi Arabia.