QXAI: Explainable AI Framework for Quantitative Analysis in Patient Monitoring Systems

Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Juan D. Velásquez, and Niall Higgins

***Abstract*—Artificial Intelligence techniques can be used to classify a patient's physical activities and predict vital signs for remote patient monitoring. Regression analysis based on non-linear models like deep learning models has limited explainability due to its black-box nature. This can require decision-makers to make blind leaps of faith based on non-linear model results, especially in healthcare applications. In non-invasive monitoring, patient data from tracking sensors and their predisposing clinical attributes act as input features for predicting future vital signs. Explaining the contributions of various features to the overall output of the monitoring application is critical for a clinician's decision-making. In this study, an Explainable AI for Quantitative analysis (QXAI) framework is proposed with post-hoc model explainability and intrinsic explainability for regression and classification tasks in a supervised learning approach. This was achieved by utilizing the Shapley values concept and incorporating attention mechanisms in deep learning models. We adopted the artificial neural networks (ANN) and attention-based Bidirectional LSTM (BiLSTM) models for the prediction of heart rate and classification of physical activities based on sensor data. The deep learning models achieved state-of-the-art results in both prediction and classification tasks. Global explanation and local explanation were conducted on input data to understand the feature contribution of various patient data. The proposed QXAI framework was evaluated using PPG-DaLiA data to predict heart rate and mobile health (MHEALTH) data to classify physical activities based on sensor data. Monte Carlo approximation was applied to the framework to overcome the time complexity and high computation power requirements required for Shapley value calculations.**

***Index Terms*—Explainability, Shapley, Attention, Monte Carlo, Vital Signs, Physical Activities**

I. Introduction

In the realm of modern healthcare, the integration of cutting-edge technology, notably through remote monitoring systems, represents a pivotal advancement in patient care and the management of diseases. These systems play an essential role in the prompt detection and averting of grave health events, chiefly through their capacity to precisely monitor and scrutinize vital signs such as temperature, pulse, respiratory rate, and mean arterial pressure [1, 2]. However, the scope of traditional monitoring systems is often constrained to displaying a patient's current health status, which limits their effectiveness in preemptively predicting and managing potential health complications.

The advent of Artificial Intelligence (AI) and deep learning heralds a new era in healthcare, transcending the boundaries of traditional methods by offering predictive insights that are indispensable for early and effective medical interventions [3, 4]. Nevertheless, these advanced methodologies come with their own set of complexities, chief among them being the lack of transparency and comprehensibility in deep learning models. These models, often labeled as ”black-box” models, pose a significant challenge in elucidating how input factors correlate with the predictive outcomes [5, 6]. This issue is particularly critical in healthcare, where understanding the rationale behind AI-driven decisions is vital for their acceptance in clinical settings and for ensuring ethical applications of such technologies.

In response to these challenges, our research presents an innovative Explainable AI framework tailored for Quantitative data (QXAI), ingeniously amalgamating the Shapley values concept [7] with an attention mechanism in the realm of deep learning models. Our approach is uniquely poised to demystify AI predictions on both granular (local) and aggregate (global) scales. It provides insightful revelations on how each individual feature contributes to specific input records and offers a comprehensive overview of feature contributions throughout the entire model. This dual-level explanation capability is adeptly employed in our framework for the purpose of predicting human vital signs and classifying physical activities, utilizing two advanced deep learning models: Artificial Neural Networks (ANN) and attention-based Bidirectional LSTM (BiLSTM). The empirical evidence from our study highlights the framework's proficiency in delivering detailed Shapley values and attention weights for each input feature, thereby clarifying their respective impacts on the outcomes of deep learning models.

Recognizing the computational demands in calculating Shapley values for extensive datasets, we have judiciously integrated the Monte Carlo method of approximation with random sampling. This strategic addition not only mitigates the computational complexities but also augments the practical utility of our framework across a spectrum of real-world applications.

Overall, our study represents a significant advancement in the field of explainable AI within healthcare. The key

Thanveer Shaik and Xiaohui Tao are with the School of Mathematics, Physics & Computing, University of Southern Queensland, Toowoomba, Queensland, Australia (e-mail: Thanveer.Shaik@usq.edu.au, Xiaohui.Tao@usq.edu.au).

Haoran Xie is with the Department of Computing and Decision Sciences, Lingnan University, Tuen Mun, Hong Kong (e-mail: hrxie@ln.edu.hk)

Lin Li is with the School of Computer and Artificial Intelligence, Wuhan University of Technology, China (e-mail: cathylilin@whut.edu.cn)

Juan D. Velásquez is with the Industrial Engineering Department at University of Chile, Chile (e-mail: jvelasqu@dii.uchile.cl)

Niall Higgins is with Metro North Hospital and Health Service, Royal Brisbane and Women's Hospital, and also with School of Nursing, Queensland University of Technology, Brisbane, Australia (e-mail: Niall.Higgins@health.qld.gov.au).

contributions of our research are:

- Development of an innovative, adaptable Explainable AI framework (QXAI) for quantitative data analysis in healthcare. This framework uniquely combines attention layer mechanisms with Shapley values within deep learning models, setting a new standard in AI explainability.
- Comprehensive evaluation of the framework's explainability capabilities, focusing on the importance of features and providing both local and global explanations. This dual approach significantly enhances the understanding of AI models, offering insights into their cognitive and behavioral aspects.
- Adoption of the Monte Carlo method to address the computational challenges in calculating Shapley values, especially for large datasets. This method significantly reduces the computational overhead, making the framework more practical for real-world applications.
- Establishment of a new paradigm in patient monitoring systems for interpreting and explaining AI predictions related to vital signs and physical activity classification. This advancement is pivotal for clinical decision-making, offering a more nuanced and in-depth understanding of patient health dynamics.

The remainder of the article is organized as follows: Section II presents related works on explainability in healthcare applications. Section III presents a formal definition of the research problem addressed. Section IV details the novel QXAI framework to explain prediction and classification problems proposed in this study. Experimental design, dataset description, data modelling, and traditional models are discussed in Section V. In Section VI, experimental results of the QXAI framework are discussed, along with its explainability and feature identification performance. Section VII discusses the random sampling approximation using the Monte Carlo method. In Section VIII, we discuss implications, strengths, and limitations of the study. Finally, the paper concludes with Section IX.

II. Related Work

In the realm of remote patient monitoring systems, the primary objective is to promptly identify high-risk patients, enabling clinicians to allocate resources effectively and intervene in a timely manner. The integration of machine learning and AI in these systems has led to significant advancements in predictive healthcare.

A. Machine Learning in Healthcare Prediction

Gong et al. [8] developed a machine learning-based framework for predicting acute kidney injury (AKI), showcasing an end-to-end decision support system that encompasses data pre-processing, risk prediction, and model explanation. This framework utilized logistic regression, random forest, and a voting-based ensemble model, along with gradient boosting algorithms, to address the challenges posed by imbalanced datasets. The model's prediction capability within 48 hours was complemented by SHapley Additive exPlanations (SHAP) values for a dual perspective: a global view highlighting critical factors and a local view detailing individual patient-level feature contributions. In addition, Wu et al. [9] compared eight feature selection methods to enhance AKI prediction, underlining the importance of feature selection stability and similarity.

B. Assessment of Interpretability Techniques

ElShawi et al. [10] proposed quantitative measures to assess the quality of several model-agnostic interpretability techniques, including LIME, SHAP, Anchors, and others. Their study utilized a random forest model to predict mortality and diabetes risk, evaluating the performance of these interpretability techniques in terms of similarity, bias detection, execution time, and trust. In a separate study, Elshawi et al. [11] applied global and local explainability techniques to predict the risk of hypertension, enhancing the transparency of machine learning outcomes. Ilic et al. [12] introduced an explainable boosted linear regression (EBLR) algorithm for time series forecasting, demonstrating that maintaining interpretability does not necessarily compromise model performance.

C. Attention Mechanism in Deep Learning

The attention mechanism, initially a breakthrough in machine translation tasks, has been adapted for healthcare applications. Bari et al. [13] conducted an empirical evaluation of attention-based deep neural networks, assessing prediction performance, explainability correctness, and sensitivity. Their results indicated that multi-variable LSTM models with explainability features performed well with complex data. Kaji et al. [14] implemented an attention-based LSTM model for predicting medical conditions like sepsis and myocardial infarction, using MIMIC-III dataset patient data. They highlighted the importance of the attention layer in extracting influential input features for better explainability. Chen et al. [15] further advanced this field by proposing bilateral asymmetric skewed Gaussian attention (bi-SGA) to improve the performance and interpretability of deep convolutional neural networks.

D. Gap in Literature and Study Contribution

The literature reveals that while deep learning is capable of predicting vital signs with minimal healthcare domain knowledge, its lack of explainability remains a significant drawback. This underscores the need for explainable AI methods to demystify the results produced by these ”black-box” models. Our study addresses this gap by introducing a novel framework that not only estimates feature importance for enhancing explainability but also provides both global and local interpretations of deep learning model predictions. This comprehensive framework aims to balance the trade-off between deep learning model performance and its explainability, thereby contributing significantly to the field of predictive healthcare.

III. Research Problem

The central research problem tackled in this study is the elucidation of deep learning model results, particularly the interpretation of predictions based on independent feature inputs in healthcare settings. This task involves comprehending the causal relationships between input factors and their effect on model predictions. It's crucial for healthcare professionals to grasp the rationale behind AI-driven predictions, understanding how variations in input feature values can influence these predictions. In a scenario where a deep learning model M uses N features, denoted as x_j (where $j = 1, \ldots, N$), to predict an output y, the research aims to elucidate how each input feature x contributes to this prediction. This understanding is vital for models where weights w_j are applied to respective features j at different layers of model M. This process can be mathematically represented as:

$$y \longleftarrow f_M(w_j \cdot x_j) \tag{1}$$

To enhance the explainability of predictions from complex, non-linear models such as neural networks and deep learning, it is essential to quantify the contribution of each feature, φxj, in a comprehensible manner. To enhance the explainability of non-linear model predictions, the contribution of each feature φ_{x_j} can be estimated into two patterns.

$$\varphi_{x_j} = w_j * x_j - E(w_j * X_j) \tag{2}$$

$$\sum_{j=1}^{N} \varphi_{x_j} = \sum_{j=1}^{N} w_j * x_j - E(w_j * x_j) \tag{3}$$

- The first pattern estimates the model output with each feature and subtracts the output with the average effect of all the features, $E(w_j * X_j)$ as shown in Equation 2. The same approach can estimate the contributions of all features. Summing up all the features' contribution in a prediction instance is, where Equation 3 shows the predicted value $f_M(x)$ minus the average predicted value $E(f_M(x))$ for the instance x.
- The second pattern adds an attention layer to the non-linear model and enables the model to focus on certain important features contributing to the output. This pattern creates a representation hj with $j = 1, \ldots, N$ of each input in vector space, and the weighted sum of the representation act as context vectors as shown in Equation 4. Extracting the weights for each input feature can influence output feature contribution φ_{x_j}.

$$c = \sum_{j=1}^{N} \alpha_j h_{x_j} \tag{4}$$

In this current study, the two patterns estimate feature contribution to explain the prediction process of the deep learning model.

IV. Explainable AI for Quantitative data (QXAI)

In this section, Explainable AI for Quantitative data (QXAI) is proposed to estimate input feature importance in deep learning model results that could be prediction or classification tasks. The proposed framework can provide explainability at two levels, one is post-hoc explainability using Shapley values and the other is intrinsic explainability using attention mechanism as shown in Fig. 1.

A. Shapley Values Calculation

To explain the contribution of input features, the Shapley value concept based on a coalition game was adopted [7]. The coalition game theory can be defined by designating a value for each coalition game with a limited set of players N, $S \subseteq N$ to be a subset of $|S|$ players and a characteristic function $v : 2^N \to \mathbb{R}$ from the set of all possible coalitions of players to a set of players that satisfies $v(\emptyset) = 0$ where $(\emptyset)$ is an empty set. This function determines each player's contribution to the outcome, and the game can be called a profit game or value game.

The profit game or value game can be adapted to the proposed QXAI framework to determine players (features) contributing to the prediction capacity of a trained deep learning model. To attribute a value to the contribution of each feature, the Shapley value concept can be adapted to explain the contribution in terms of expected marginal contribution. Shapley values assume that all the features contribute to the outcome, and the amount that each feature x_j contributes in a coalition game (v, N) is shown in Equation 5.

$$\varphi_{x_j}(v) = \sum_{S \subseteq N \setminus \{x_j\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{x_j\}) - v(S)) \tag{5}$$

where the sum extends over all subsets S of N not containing feature i and n is the total number of features.

The above Equation 5 can further break-down to have individual feature contribution as $v(S \cup x_j) - v(S)$. The characteristic function $v(S)$ can be calculated by using Kernel SHAP.

$$\varphi_{x_j}(v) = \frac{1}{n!} \sum_{R} \left[v(P_{x_j}^R \cup \{x_j\}) - v(P_{x_j}^R) \right] \tag{6}$$

where the sum iterates over all $n!$ orders R of the features and $P_{x_j}^R$ is the set of features in N which proceeds the order R.

In simple terms, Shapley of a feature x_j can be defined as below, Equation 7:

$$\varphi_{x_j}(v) = \frac{1}{n} \sum_{K} \frac{\varphi(x_j)}{Z} \tag{7}$$

Where n is a number of features, $\varphi(x_j)$ is marginal contribution of feature x_j to coalition, K is coalitions excluding x_j, Z is a number of coalitions excluding x_j.

Shapley proposed four conditions (or axioms) below that must be satisfied to have fair contribution of features to a prediction. Equations 5,6 obey these conditions while estimating the contribution value of each feature.

- The summation of Shapley values of all agents equals the value of the total coalition.
- All features have a fair chance to participate in a prediction outcome by including in all permutations and combinations of the features.
- If a participated feature x_j contributes nothing to a prediction outcome, then zero value is attributed to the feature's contribution.

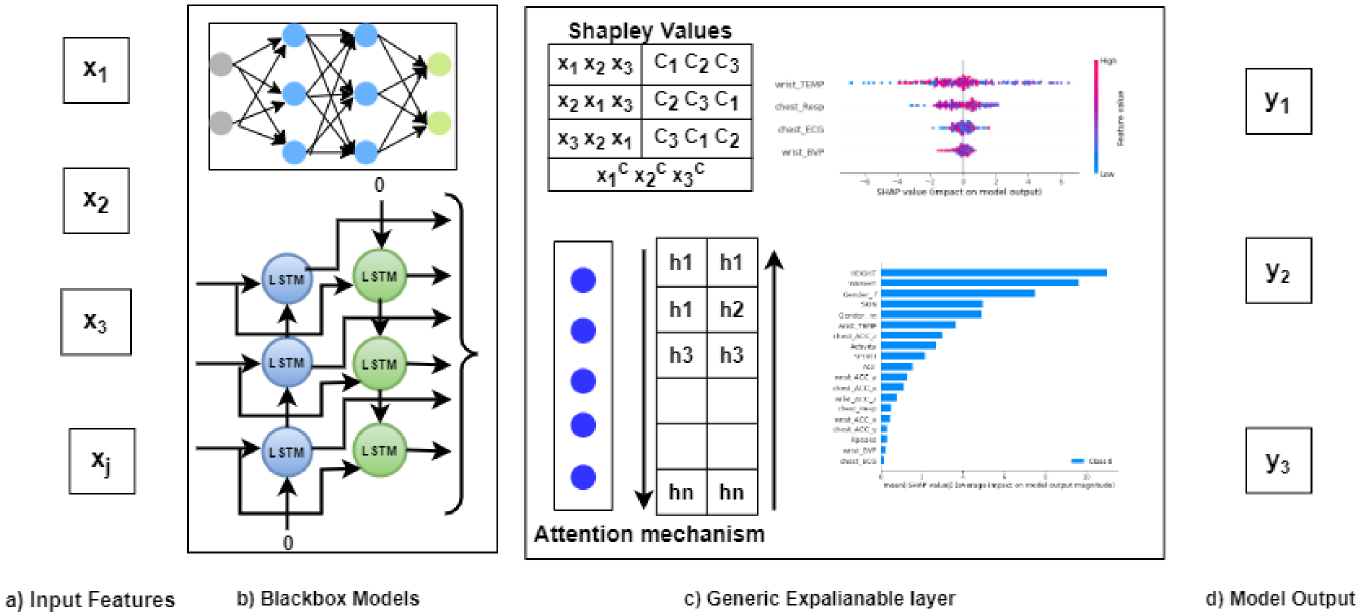


Fig. 1: Explainable AI (QXAI) framework

- For any pair of predictions $v, w : \varphi(v+w) = \varphi(v) + \varphi(w)$ in which the values are based on additive property $(v+w) = v(S) + w(S)$ for all subsets S.

B. Kernel SHAP

Kernel SHAP is a model-agnostic method from the combination of classical Shapley values discussed in Equations 5,6 and local explainable model-agnostic explanations (LIME) to approximate SHAP values. Instead of retraining models with a subset of features $|S|$, the full model f can be used which is already trained, while replacing missing features with marginalized features. Considering an instance with three features x_1, x_2, x_3 and following Equation 8 estimates a partial model with x_3 being missed. However, $p(x_3)$ is still required to approximate the missing x_3 feature. To address this, a custom proximity function π from LIME as shown in Equation 9 and SHAP similarity kernel equation 10 can be used.

$$f_{x_1,x_2}(x_1,x_2) \longrightarrow \int f(x_1,x_2,x_3)p(x_3)dx_3 \tag{8}$$

$$\pi_x^{LIME}(z) = \exp(-D(x,z)^2/\sigma^2) \tag{9}$$

$$\pi_x^{SHAP}(z') = \frac{(p-1)}{\binom{p}{|z'|}|z'|(p-|z'|)} \tag{10}$$

Equation 9 penalizes the distance between sample points and the original features' data, for which explainability is being estimated. In Equation 10, coalitions with a number of features that are far from 0 and p will be penalized. The equation adds more weight to coalitions with a small set of features or almost all the features to highlight the independent behavior of the feature or the impact of the features in interaction with others. The choice of this SHAP similarity kernel is based on three properties of additive feature attribution methods local accuracy, missingness, and consistency [7]. In this study, Kernel SHAP is used to estimate the contributions of each feature x_j value to the prediction. It consists of five steps: 1) Sample coalitions with features and without features. 2) Get prediction for each sample coalition by first converting to the original feature space and applying the machine learning model. 3) Estimate the weight for each coalition with the SHAP kernel. 4) Fit the weighted linear model. 5) Return Shapley value $\varphi_{x_j}(v)$ and the coefficients of the model.

C. Attention Mechanism

The attention mechanism is a widely adopted concept in Natural Language Processing (NLP) tasks like neural machine translations and extracting the cause-effect of input features to model output [16, 17]. The attention mechanism predicts the outcome with better accuracy because its cognitive capability can enhance certain parts of important input data for deep learning model training. The idea of using the attention mechanism to model explainability is to identify the weights beings assigned to each input feature in predicting the outcome. This assists in decoding the importance of each feature and enables human explanation of the cause-effect of the input features.

An attention layer added to a deep learning model can mimic the cognitive capability of the attention mechanism. Given a set of input features N, x_j is a feature value, with $j = 1, \ldots, N$ to predict an output value y. A Bidirectional Long Short-Term Memory (BiLSTM) model can generate vector representations of the input features, such as h_j with, $j = 1, \ldots, N$ based on the forward and backward hidden states in the deep learning model. A generic encoder-decoder model focuses on the last state of the encoder LSTM model and uses it as a context vector. This would cost the information loss of previous states. Attention acts as an interface between the encoder and decoder states of the BiLSTM model and provides a context vector to the decoder with information from every encoder's hidden states. For each prediction value y, a context vector c is generated using the weighted sum of the vector representations, as shown in Equation 11. The weights α_j are computed using a softmax function as shown in Equation 12. The output score e_j is calculated in a feedforward neural network described by a function f to capture alignment between input feature x_j and output y. The input features are then multiplied (dot product) with (w_j+B) where w_j is weight and B bias followed by a tan hyperbolic function to estimate the score e_j as shown in Equation 13

$$c = \sum_{j=1}^{N} \alpha_j h_{x_j} \tag{11}$$

$$\alpha_j = softmax(e_j) = \frac{exp(e_j)}{\sum_{j=1}^{N} exp(e_j)} \tag{12}$$

$$e_j = f(x_j, h_{x_j}) = tanh(xj \cdot (w_j + B)) \tag{13}$$

For input features x_1, x_2, x_3, x_4 , let the weights α_j be, $[0.2, 0.4, 0.6, 0.1]$ then the context vector would be as shown in equation 14. This can assist in estimating the importance of each input feature in the context vector, which will be fed to the decoder network for model predictions.

$$c = 0.2 \times x_1 + 0.4 \times x_2 + 0.6 \times x_3 + 0.1 \times x_4 \tag{14}$$

D. Global and Local explanation

Two different forms of explanation perspectives such as global explanation and local explanation are proposed in this study. The global explanation can provide the contribution of each feature in the prediction of vital sign. This is designed to assist clinicians by providing holistic information about the prediction and to identify which clinical factors or features need special attention. To estimate the global importance of the features in the prediction, the absolute Shapley values calculated from Equation 5 are averaged for each feature across the data, as shown in Equation 15. Based on this calculation, the features can have their importance sorted in descending order.

$$I_i = \frac{1}{n} \sum_{i=1}^{n} |\varphi_i| \tag{15}$$

Although feature importance can provide an overview of all selected features' importance towards a prediction, it cannot uncover the correlation of the features with a target variable and estimate contributing and non-contributing data points of a feature. This, however, can be achieved by using Shapley values of each feature on a summary plot showing the level of positive and negative contribution to a target variable.

In the case of local explanation, vital signs prediction at each time step can be decrypted. This can summarize features that are aiding the patient's health in terms of vital signs and can enable personalized monitoring, which is critical in healthcare applications. The Shapley values of each feature can be positive or negative, and each value is considered a force that either increases or decreases the prediction value. This helps to explain individual features that are forcing the prediction value to either increase or decrease. The local explanation concept can be applied to an individual record in a prediction or a group of records related to a specific subject or activity.

Algorithm 1 Feature contribution estimation

```
Require: a set of features $\mathcal{F} = \{1, 2, \ldots, N\}$;a set of deep
    learning models $\mathcal{M} = \{m_1, m_2\}$ where $m_1$ is without
    attention and $m_2$ is with attention;a input dataset D
Ensure: Contributions of the features $\mathcal{F} = \{1, 2, \ldots, N\}$ in
    the form of Shapley values and attention weights;
 1: Split dataset: $D = D^{train} \vee D^{test}$
    Global explanation
 2: $m_1^{train} \longleftarrow D^{train}$
 3: $m_1^{test} \longleftarrow D^{test}$
 4: $Shapley\_values \leftarrow kernelshap(m_1^{train}, D^{test})$
 5: $m_2^{train} \longleftarrow D^{train}$
 6: $m_2^{test} \longleftarrow D^{test}$
 7: $attention\_weights \leftarrow model.attention\_weights()$
    Local explanation
 8: for d in D do
 9:     $Shapley\_values \leftarrow kernelshap(m_1^{train}, d)$
10:     $attention\_weights \leftarrow m_2.attention\_weights()$
11: end for
```

E. QXAI Algorithm

The proposed QXAI framework comprises two deep learning model approaches, one with model attention and the second without. The framework can be implemented with the Algorithm 1 and can be adapted to execute global and local explanations. In Algorithm 1, line 1 splits the input data into test and train sets to train and evaluate the deep learning models. Lines 2-7 present the global explanation using kernel SHAP and attention layer weights. Lines 2-4 train a deep learning model without an attention layer and pass it to the kernel SHAP explainer to extract Shapley values of the input features. Lines 5-7 present the attention-based deep learning model and extracts the attention layer weights, thus defining input feature importance. Lines 8-11 present the local explanation for each input record d from data D.

V. Experiment

The two key aspects of an explainable AI framework are the understanding phase and the explaining phase [18]. The former is concerned with improving models during training by interpreting critical features and building robust models, while the latter involves deploying and providing human-readable explanations to end users. Striking a balance between model performance and explainability is always a challenge in AI applications. In AI applications, there is always a trade-off between model performance and explainability [19]. According to Zacharias et al. [20], the preprocessing stage, specifically feature selection, has been overlooked in explainable AI applications and requires attention. The importance of each feature to the outcome can be used for semantic labeling

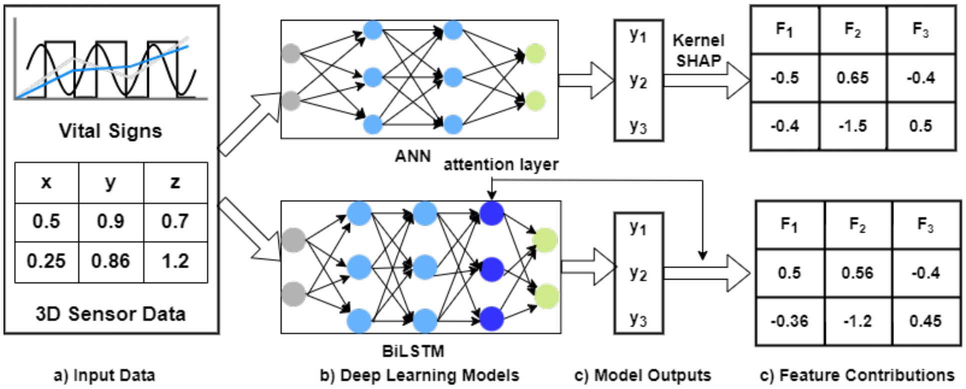


Fig. 2: Experimental Design

and to improve cognitive understanding, as it provides positive framing and direction (positive or negative contribution).

To address the limitations of explainable AI, the proposed QXAI framework in this study focuses on feature selection and provides local and global explanations through post-hoc models and intrinsic weights. The study evaluates feature importance in the QXAI framework, which can reduce dimensionality, improve cognitive understanding, and help with decision-making. Good explanations are crucial for making informed decisions, especially in dynamic domains like healthcare. In addition to the feature importance step in explainability, this study further breaks down the explainability into local and global explanations for each supervised learning task in classification and regression. Local and global explanations help in understanding the positive or negative contribution of features at the model and individual levels. The study used publicly available benchmark datasets for evaluation. Figure 2 illustrates the experimental design of the proposed framework.

A. Datasets

- **PPG-DaLiA** [21]: This dataset from 15 subjects comprised physiological and motion data while performing a wide range of activities under close to real-life conditions. The collected data were from both a wrist-worn (Empatica E4) and a chest-worn (RespiBAN) device. The dataset consists of 11 attributes, including 3-dimensional (3D) acceleration data, electrocardiogram (ECG), respiration, blood volume pulse (BVP), electrothermal activity (EDA), and body temperature.
- **MHEALTH** [22]: This dataset comprises the body motion of ten volunteers while performing 12 physical activities recorded from three sensors at the chest, left ankle, and right lower arm. There were 21 independent attributes including acceleration, gyroscope, and magnetometer of the three sensors. A dependent variable classifying the 12 activities was based on the sensor data.

B. Data Modelling

Datasets consisted of preprocessed raw data from the sensor's signal and features were stored in different CSV files. In this step of data preparation, the dataset was further preprocessed to have a single structured file with a set of features for each subject. The datasets were prepared for two different tasks: regression and classification. The regression task was to predict the heart rate of the subjects based on their sensor readings. The classification task was to classify the physical activities of the subjects based on their motion data recorded from three axes of sensors. The physical activities label was preprocessed to have a multi-label classification. Each of these datasets was split into an 80:20 ratio for 80% of data for training and 20% of data for testing.

In this study, two deep learning models artificial neural networks (ANN), and Bidirectional LSTM (BiLSTM) models were adopted. The ANN model was configured with an input layer, hidden layers, and an output layer. The traditional activation function rectified linear unit (ReLU) has a limitation of defining negative inputs to zero which deactivates the nodes or neurons. Considering the negative values in 3D sensor data, the ANN model used the activation function LeakyReLU in input and hidden layers to avoid the zero input values of the negative attributes. The output layer was configured with the traditional activation function ReLU to predict the target variable heart rate greater than zero based on the activation function property. The loss function used for the regression study was mean absolute error, which also acted as a performance metric for the model. For the classification task, binary cross entropy acted as a loss function along with metrics like accuracy. The Adam adaptive optimizer [23] was chosen for the model for its quick computational time, it requires fewer parameters for tuning compared to other optimizers. The attention mechanism discussed in the proposed framework was added to the BiLSTM model, which has encoder and decoder states to generate vector representations. The preprocessed data was fed to the attention-based BiLSTM model and extracted the attention layer weights. This determined the input feature importance in the deep learning model prediction.

The datasets in this study were created by preprocessing raw data from sensor signals and storing the features in separate CSV files. These datasets were then combined into a single structured file for each subject, with separate datasets prepared for regression and classification tasks. The regression task involved predicting the subject's heart rate based on sensor

TABLE I: Implementation details

	Regression		**Classification**	
	Shapley Values	**Attention Mechanism**	**Shapley Values**	**Attention Mechanism**
Models	ANN	BiLSTM	ANN	BiLSTM
No of Layers	5	4+attention layer	5	4+attention layer
Activation Functions	relu, sigmoid	relu, Softmax	LeakyReLU, Sigmoid	relu, Softmax
Optimizers	Adam		Adam	
loss Functions	mean_absolute_error		binary_crossentropy	
Epochs	100		100	
Batch Size	64		64	

readings, while the classification task involved categorizing the subject's physical activities using motion data from three axes of sensors. The datasets were split into 80% for training and 20% for testing. Two deep learning models, ANN and BiLSTM, were used in this study as shown in Tab. I. The table presents implementation details of ANN and BiLSTM models in regression and classification tasks.

For regression tasks, the models use the Shapley Values and attention mechanism. The ANN model has 5 layers with the activation functions of relu and sigmoid. The BiLSTM model has 4 layers with an additional attention layer with the activation functions of ReLU and softmax. The optimizer used is Adam, and the loss function is mean_absolute_error. For classification tasks, the models also use ANN and BiLSTM architectures, Shapley Values, and attention mechanisms. The ANN model has 5 layers with the activation functions of LeakyReLU and sigmoid. The optimizer used is Adam, and the loss function is binary_crossentropy. The BiLSTM model has 4 layers with an additional attention layer. The activation functions used are ReLU and softmax. For both prediction and classification tasks, the models are trained for 100 epochs with a batch size of 64.

C. Traditional Models

By comparing the feature importance estimated using Shapley values and intrinsic weights of the attention mechanism with the traditional machine learning models, the explainability of the proposed framework was evaluated. The two deep learning models in the framework, ANN and BiLSTM, were also evaluated to ensure high performance and robustness with explainability. This allowed the study to evaluate the effectiveness of the framework in explainability without compromising model performance.

The proposed approach was evaluated with models with state-of-art performances. The deep learning models adopted in the proposed approach were compared with heart rate prediction and human activity recognition performances. The feature importance was compared with traditional machine learning models, which had the capability to produce feature importance for prediction and classification results.

Prediction

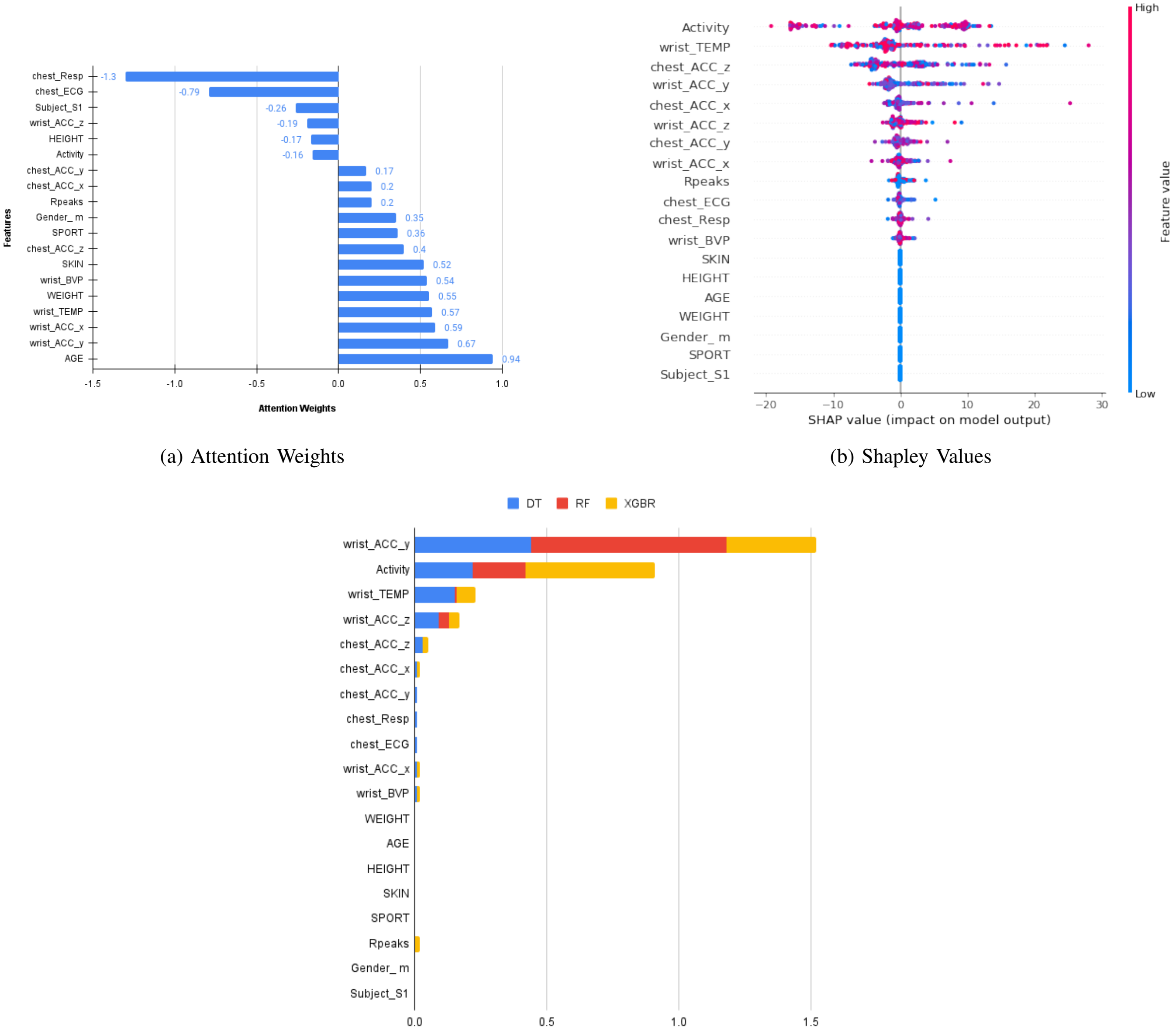
- Ni et al. [24] proposed context-aware sequential models to capture personalized fitness data and forecast heart rate to recommend suitable activities. The authors used a multi-layer perceptron model to forecast heart rate.
- Zhu et al. [25] proposed four LSTM models for an optimization training system to predict heart rate under three different types of exercises walking, rope jumping, and running. Three of the four LSTM models were used for heart rate prediction and one for human activity recognition.

Classification

- In a previous study, we proposed FedStack [26], a novel federated framework to classify patients' physical activities. We adopted deep learning models such as CNN, ANN, and BiLSTM for the classification.
- Bozkurt et al. [27] compared deep learning model performance with traditional machine learning models for human activity recognition. Deep Neural Network (DNN) model achieved an accuracy of 96.81% and outperformed other models.

Feature Importance

- Li et al. [28] proposed an explainable machine learning model named cardiac arrest prediction index for early detection of cardiac arrest. The authors used the XGBoost model for the prediction and achieved an area under the receiver operating characteristic curve (AUROC) of 0.94.
- Gong et al. [8] used XGBoost and voting ensemble method combining random forest and logistic regression to predict acute kidney injury. For explanation, the SHAP technique was used to understand important predictors and relationships among the predictors.
- Ali et al. [29] proposed supervised machine learning algorithms such as Random Forest, Decision Tree, and KNN for heart disease prediction. Feature importance scores for each feature were computed with Decision Tree and Random Forest [30].



(a) Attention Weights

(b) Shapley Values

(c) Traditional Models

Fig. 3: Regression Model—Feature Importance Plots

D. Performance Metrics

Explainability is a multifaceted concept, and there is no single metric to measure it. The evaluation of explainability involves comparing the feature importance provided by different models, such as comparing the explanations of ANN and BiLSTM with those of traditional models. In this study, another two sets of performance metrics were adapted to evaluate deep learning models' prediction and classification results. For prediction, mean absolute error (MAE) and mean squared error (MSE) was used to evaluate the performance of the prediction model. Both metrics measure the deviation or difference of a predicted value from its actual value. For classification, a traditional confusion matrix was used to calculate precision, F1-Score, recall, and balanced accuracy metrics of deep learning results on multi-label classification.

VI. Experimental Results and Analysis

In this section, we analyze the evaluation results of the proposed QXAI framework. The results are focused on explainability in terms of feature importance for positive framing, local explanations for semantic labelling to explain the positive or negative contributions of each input feature to the deep learning model's prediction, and global explanations that can explain a model's overall predictions with interactive plots. To address the trade-off between explainability and model performance in AI applications [31], the performance of the deep learning models, ANN and BiLSTM-attn, in the framework for both regression and classification tasks was evaluated and compared with those of traditional machine learning models.

TABLE II: QXAI Prediction Performance

Model	MAE	MSE
ANN	**3.33**	**24.51**
BiLSTM-attn	**4.40**	**43.72**
MLP [24]	4.71	47.95
LSTM [25]	5.54	69.03

A. *QXAI in Regression Problem*

The proposed QXAI approach was evaluated on its ability to predict heart rate based on sensor data and clinical indicators. Other vital signs retrieved from human subjects were in the PPG-DaLiA dataset. The two deep learning models ANN and attention-based BiLSTM proposed in the framework were trained on the data to predict the vital signs. The models' performance was compared with other traditional models shown in Tab. II. The ANN model performed better than the attention-based BiLSTM, MLP, and LSTM models with MAE and MSE of 3.33 and 24.51 respectively.

1) Feature Importance: Feature importance of input features was estimated using the proposed QXAI approach and compared with traditional machine learning model feature importance. The three feature importance plots shown in Fig. 3a, 3b,and 3c present attention weights retrieved from the BiLSTM model, Shapley values estimated from Kernel SHAP, and traditional model feature importance respectively. The y-axes in each subplot hold the input features, with x-axes showing the importance of each feature to the respective model's prediction. The large value of the x-axis determines the importance or contribution of a feature to model performance in predicting heart rate. Activity, chest, and wrist sensors data had high feature importance for heart rate prediction compared to other input vital signs like wrist_BVP, chest_Resp, and chest_ECG. The Shapley values plot 3b and attention weights plot 3a presented the negative dimensions of each feature's contribution.

2) Explainability: As discussed in Section IV, global and local explanations both contribute to presenting a patient's health status at different levels. The local explanation assists the clinician to explain the health status at a particular time step of patient monitoring. Fig. 4a presents feature contribution to the ANN model label prediction for a selected random record. The randomly selected record is of a male subject aged 25 years, height 168 centimeters, weight 57 kilograms with fitness level 5 on a scale 1-6 where 1 refers to them exercising less than once a month and 6 refers to 5-7 times a week. The subject's activity was measured during his lunch break, and his heart rate prediction was 71.24. The red highlighted features in Fig. 4a indicated a negative contribution and pushed the prediction value to the right (higher) side of the scale, whereas the blue features positively contributed and pushed the prediction value to the left (lower) side of the scale. This infers activity, wrist_ACC_y, and wrist_ACC_x features are negatively contributing and trying to decrease the heart rate value. The Rpeaks and wrist_TEMP features are balanced by increasing the heart rate to the expected value of 72.95. The SHAP values of each feature can be positive or negative. Similarly, Fig. 4b presents a subject-level explanation of features' contribution to their heart rate prediction based on 200 records. The chart is related to a subject and presents each predicted value on the y-axis with its feature contribution spread on the x-axis in blue and red highlight. This is an interactive plot with dropdowns on the x-axis and y-axis changing and shows the impact of individual features on all 200 predictions. The plot is a screenshot of a prediction value of 107.9 in which the feature activity from wrist_ACC_x and wrist_TEMP are negatively contributing to the heart rate prediction.

TABLE III: QXAI Classification Performance

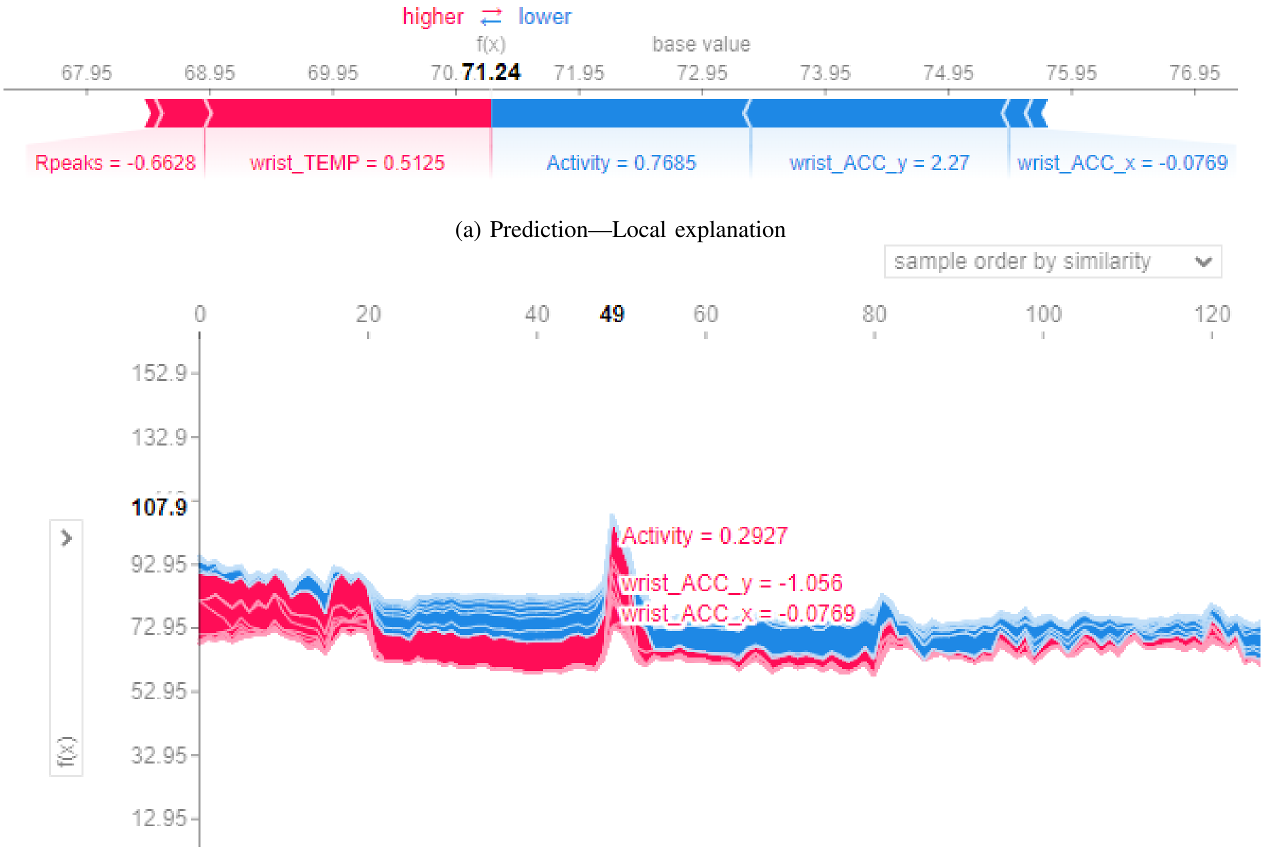
	Precision	Recall	F1-score	Balanced Accuracy
ANN	1	1	1	1
BiLSTM-Atten	0.92	0.78	0.77	0.88
CNN [26]	0.99	0.98	0.98	0.98
DNN [27]	0.97	0.97	0.97	0.97

B. *QXAI in Classification Problem*

The proposed QXAI approach was also used to explain the classification of human physical activities. Both the deep learning models ANN and attention-based BiLSTM were trained on the MHEALTH dataset. Model classification performance was compared to DNN and CNN, as shown in Tab. III. The ANN model had the best performance, with all evaluation metric values equalling 100%. CNN and DNN models also performed better than the attention-based BiLSTM model. The proposed framework disclosed the intrinsic weights of each feature in classification and post-hoc model explanations with Shapley values.

1) Feature Importance: The Shapley values and attention weights computed from the deep learning models determined the input feature importance in classifying human physical activities. Feature importance from the deep learning model was compared with feature importance in traditional machine learning models as shown in Fig. 5. The y-axes in all three subplots, 5a, 5b,and 5c refer to the 21 input features passed to the deep learning and the x-axes present the importance of a feature to model classification results. The attention-based BiLSTM model assigned more negative weights to all the input features. The sensor attributes at the wrist and ankle area were assigned with more weights in terms of magnitude to classify human physical activities as shown in Fig. 5a. The Shapley values plot 5b shows full body motion activities such as climbing stairs, jogging, walking, running, and jump front & back rely on left ankle sensor gyroscope data. The feature importance metrics from traditional machine learning models could not differentiate labels in their plot, as shown in Fig. 5c, but the results show that gyroscope data features contribute more to physical activity classification.

2) Explainability: The patients' physical activity classification can be explained in detail by breaking down the Shapley values with force plots as shown in Fig. 6a, 6b. The local explanation at each input record level can assist clinicians to explain physical activity classification and can explain which



(a) Prediction—Local explanation

(b) Prediction—Global explanation

Fig. 4: Explanations for prediction: (a) Local explanation illustrating individual feature contributions. (b) Global explanation showing overall feature contributions.

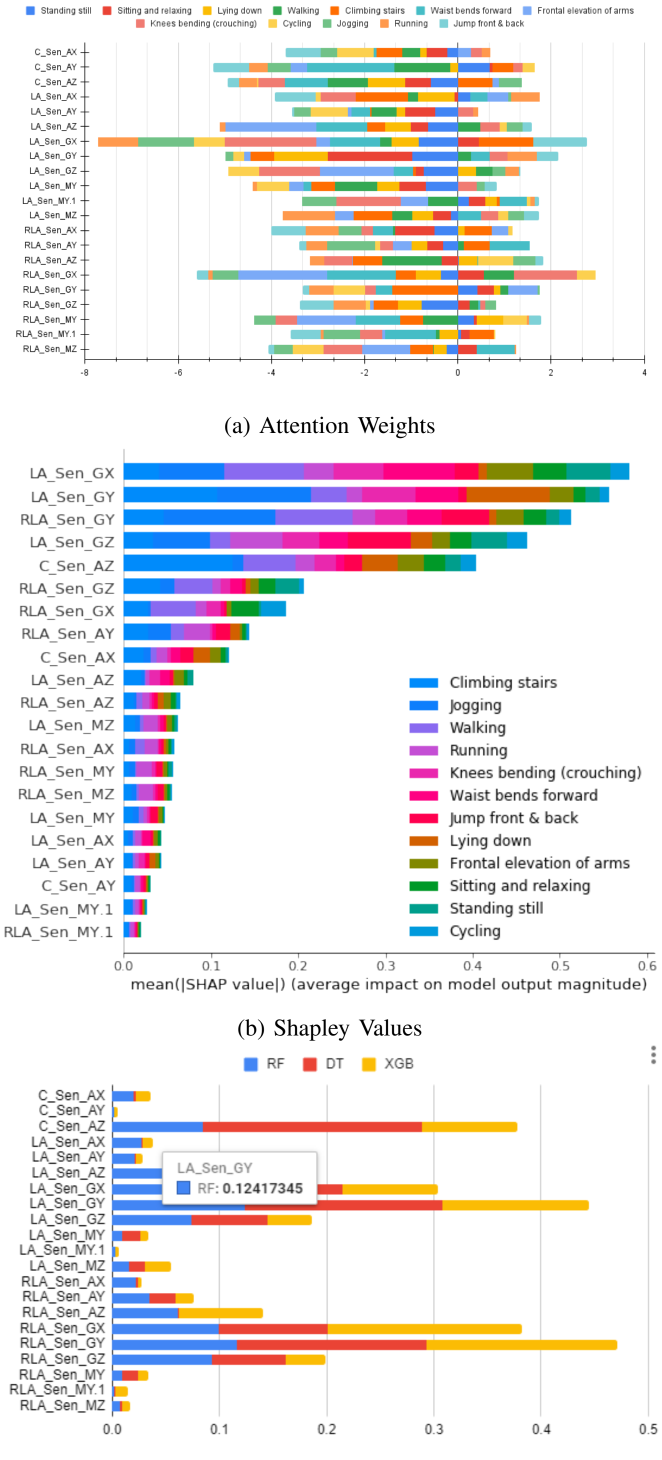
sensor features are actually contributing to the classification. In plot 6a, the ANN model prediction probability of an arbitrarily selected record shows that the x, and z dimensions of the left ankle gyroscope try to push the model probability higher but the y-axis of the right lower arm and left ankle sensors are pushing the probability negatively according to Shapley values' feature importance. Similarly, Fig. 6b presents a subject-level interpretation of features that contribute to their physical activity classification based on 200 records. The chart is related to a subject and presents each predicted value on the y-axis with its feature contribution spread on the x-axis in blue and red highlights. This is an interactive plot with dropdowns on the x-axis and y-axis to change to see the impact of the individual feature on all 200 predictions. The plot is a screenshot of a predicted value 1 in which chest sensor acceleration positively contributes and left ankle and right lower arm sensor features negatively contribute to the heart rate prediction.

VII. Monte Carlo Approximation

Feature contributions in model prediction can be estimated based on Shapley value computed using Equation 5 proposed in Section IV. These computations have an exponential time complexity and increase in number of features makes the Shapley value calculation unfeasible. In this study, Monte

Algorithm 2 Monte Carlo Approximation on Feature contribution estimation

```
Require: a set of features $x_j = \{1, 2, \ldots, N\}$;a set of deep
    learning models $\mathcal{M} = \{m_1, m_2\}$ where $m_1$ is without
    attention and $m_2$ is with attention;input data $\mathcal{D}$
Ensure: Contribution of the features $x_j = \{1, 2, \ldots, N\}$
marginal contribution $\phi_{x_j} \leftarrow \emptyset$
for all $x_j = \{1, 2, \ldots, N\}$ do
    z← random sample from $\mathcal{D}$
    x← random sample from $N$
    choose random permutation o of the feature $x_j$
    $x : x_o = x_1, \ldots, x_j$
    $z : z_o = z_1, \ldots, z_j$
    Build two new samples
    with factor $F$:
    $x_{+j} = (x_1, \ldots, x_{j-1}, z_o = z_1, \ldots, z_{j-1})$
    without factor $F$:
    $x_{-j} = (x_1, \ldots, x_{j+1}, z_o = z_1, \ldots, z_{j+1})$
    Compute marginal contribution of feature $F$:
    $\phi_{x_j} \leftarrow m_1(x_{+j}) - m_1(x_{-j})$
end for
$\hat{\phi_{x_j}} \leftarrow \frac{1}{x_j} \sum_{m=1}^{x_j} \phi_{x_j}$
```



(a) Attention Weights

(b) Shapley Values

(c) Traditional Models

Fig. 5: Classification Model—Feature Importance Plots

Carlo approximation was adopted to calculate each feature contribution as shown in Equation 16. This approximation technique can extract Shapley values for each feature for both deep learning models. The results have been discussed in this section.

$$\varphi_i = \frac{1}{N}\sum_{n=1}^{N}(f(d^m_{+i}) - f(d^m_{-i})) \tag{16}$$

where $f(\cdot)$ is the contribution of subset features. The d^m_{+i} and d^m_{-i} is the subset of with and without factor i in subset n features, respectively.

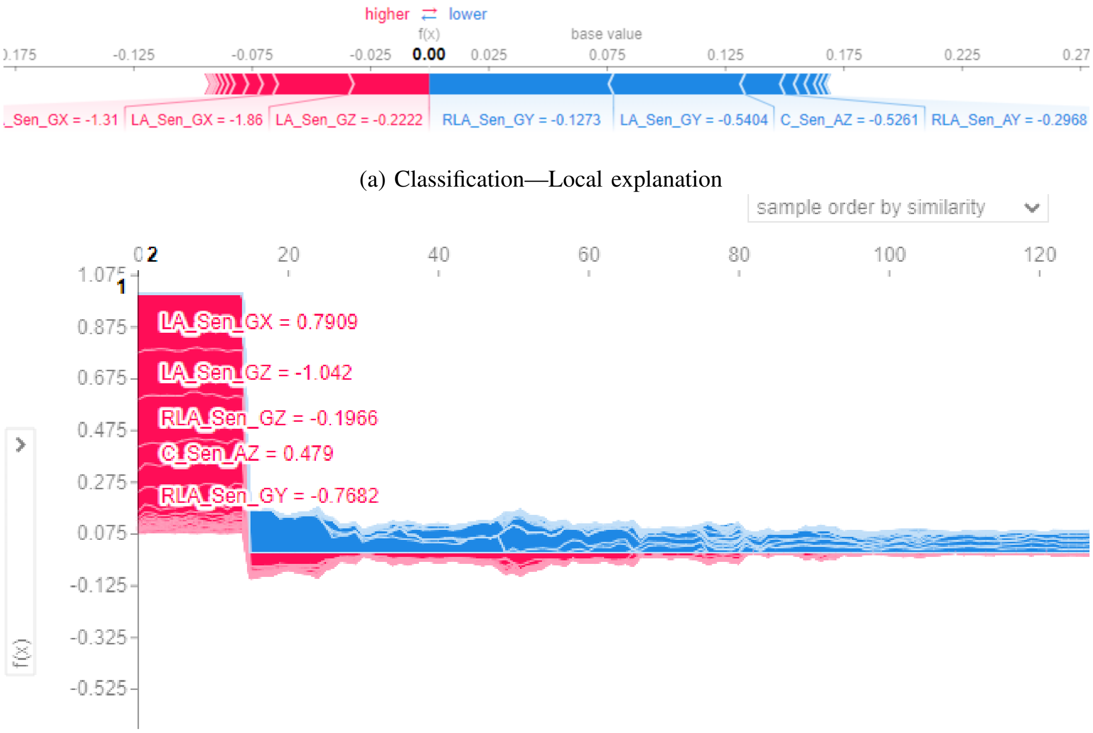
The implementation of the Monte Carlo approximation is presented in algorithm 2. Lines 3-7 obtain sampled data from the input data D. Lines 8-11 build new samples with or without consideration of a feature x_j. Line 12 calculates the marginal contribution ϕ_{x_j} of feature x_j. Lines 2-13 are a loop iterating to calculate the contribution of each feature one by one. Finally, line 14 calculates the Shapley value by averaging the outputs of multiple runs.

Monte Carlo approximation was applied to both deep learning models in the proposed QXAI framework. In ANN model prediction results, the approximation technique estimate wrist_TEMP is the most contributing feature in terms of magnitude, but the negative value shows that the feature is inversely proportional to the heart rate prediction. The other chest_ACC_z, chest_ACC_x, and chest_Resp features contributed positively towards the heart rate prediction. The attention weights from the BiLSTM model show that most of the input features are inversely proportional to the model output with negative values. The heat map shows that wrist_TEMP, wrist_ACC_z, and chest_ACC_z are the most contributing features to the heart rate prediction as shown in Fig. 7a. Similarly, the Monte Carlo approximation was applied to the deep learning classification models. The 3D axes of the sensor inputs were merged to have chest sensor acceleration, left ankle sensors' acceleration, gyroscope, and magnetometer, and right lower arm sensors' acceleration, gyroscope, and magnetometer as shown in Fig. 7b. The figure shows ANN model classification Shapley values for the consolidated input feature in the top heat map. The bottom heat map shows the attention-based BiLSTM model Shapley values. The full body activity like climbing stairs classification was more contributed by gyroscope data of the left ankle and right lower arm sensors and acceleration data of the chest sensor.

VIII. Discussion

The research presented in this paper makes a significant contribution to the emerging field of explainable AI (XAI) in healthcare, particularly by addressing the challenge of interpretability in deep learning models for vital sign prediction and physical activity classification. The proposed Explainable AI for Quantitative data (QXAI) framework is noteworthy for its innovative approach that combines Shapley values and attention mechanisms, offering a comprehensive dual perspective on both post-hoc and intrinsic explainability. This discussion delves into the implications, strengths, limitations, and future directions of this study.

Implications and Contributions: The QXAI framework addresses a critical gap in healthcare AI by providing a solution to the 'black-box' nature of deep learning models. This is crucial as the explainability of AI models is increasingly becoming a requirement, especially in high-stakes fields like



(a) Classification—Local explanation

(b) Classification—Global explanation

Fig. 6: Explanations for classification: (a) Local explanation illustrating individual feature contributions. (b) Global explanation showing overall feature contributions.

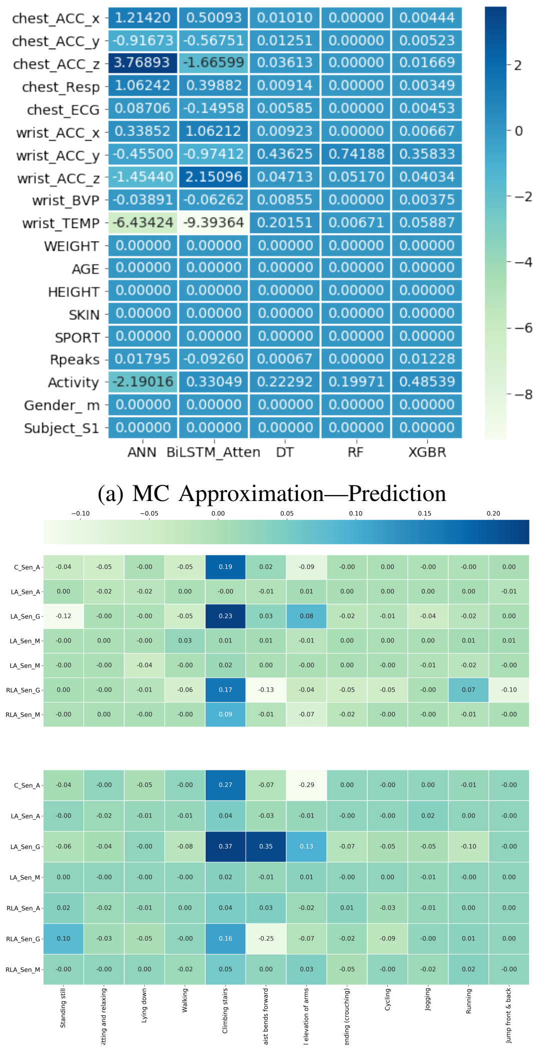
healthcare. The integration of Shapley values for post-hoc explainability and attention mechanisms for intrinsic understanding allows for a nuanced interpretation of AI decisions. This dual perspective of explainability not only enhances the trustworthiness of AI models but also makes them more practical and useful for clinicians. By enabling healthcare professionals to understand the reasoning behind AI-driven predictions, the framework facilitates informed decision-making in patient care.

Strengths of the Study: One of the major strengths of this study is the robust performance of the QXAI framework in both vital sign prediction and physical activity classification tasks. The superior performance, as compared to traditional models, highlights the potential of deep learning in enhancing healthcare diagnostics and monitoring. Furthermore, the comprehensive nature of the explainability approach employed in this study marks a significant advancement over existing methods that typically focus on either post-hoc or intrinsic explainability. The practical application of the framework, demonstrated through its effectiveness on real-world datasets like PPG-DaLiA and MHEALTH, underscores its potential for implementation in real healthcare settings.

Limitations and Future Directions: Despite its strengths, the study is not without limitations. The computational demands, particularly with large datasets due to the use of kernel SHAP, highlight the need for more efficient XAI algorithms. Additionally, while the framework shows promise, its generalizability across a broader range of healthcare scenarios remains to be tested. Future research should aim at scaling the framework for different types of healthcare data and conditions. Another area for future improvement is the user-centric design of the framework. Tailoring explanations to be intuitive for healthcare practitioners, with varying levels of technical expertise, could enhance its clinical adoption. Moreover, the integration of the framework within existing clinical workflows and ensuring data privacy and ethical AI use are crucial considerations for future development.

IX. Conclusion

In healthcare applications, the explainability of machine learning model predictions or results is critical. This can assist clinicians in understanding the results to assist with clinical decisions that take appropriate steps for treatment. Existing deep learning models have a limitation in the explainability or interpretability of their results. The prediction or classification capacity of the proposed QXAI framework is outstanding compared to traditional machine learning models, with minimal knowledge of the healthcare domain knowledge to address the research problem. To utilize the advantage of the prediction capacity, this study proposed to adopt the Shapley values concept to vital signs prediction and decode global explanation at the overall population and local explanation at the subject level. However, the study was limited by the kernel SHAP method, which required significant memory and storage for large datasets. Future directions include incorporating more diverse feature inputs to enhance remote monitoring systems for clinical decision support.



(a) MC Approximation—Prediction

(b) MC Approximation—Classification

Fig. 7: Monte Carlo Approximation for Feature Importance Analysis in Prediction and Classification.

References

[1] L. P. Malasinghe, N. Ramzan, and K. Dahal, "Remote patient monitoring: a comprehensive study," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 1, pp. 57–76, 2019.

[2] S. B. Asiimwe, E. Vittinghoff, and M. Whooley, "Vital signs data and probability of hospitalization, transfer to another facility, or emergency department death among adults presenting for medical illnesses to the emergency department at a large urban hospital in the united states," *The Journal of Emergency Medicine*, vol. 58, pp. 570–580, apr 2020.

[3] X. Tao and J. D. Velasquez, "Multi-source information fusion for smart health with artificial intelligence," 2022.

[4] N. Prakash, A. Manconi, and S. Loew, "Mapping landslides on eo data: Performance of deep learning models vs. traditional machine learning models," *Remote Sensing*, vol. 12, no. 3, p. 346, 2020.

[5] S. M. Muddamsetty, M. N. Jahromi, A. E. Ciontos, L. M. Fenoy, and T. B. Moeslund, "Visual explanation of black-box model: similarity difference and uniqueness (sidu) method," *Pattern recognition*, vol. 127, p. 108604, 2022.

[6] A. Adadi and M. Berrada, "Explainable AI for healthcare: From black box to interpretable models," in *Embedded Systems and Artificial Intelligence*, pp. 327–337, Springer Singapore, 2020.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[8] K. Gong, H. K. Lee, K. Yu, X. Xie, and J. Li, "A prediction and interpretation framework of acute kidney injury in critical care," *Journal of Biomedical Informatics*, vol. 113, p. 103653, Jan. 2021.

[9] L. Wu, Y. Hu, X. Liu, X. Zhang, W. Chen, A. S. L. Yu, J. A. Kellum, L. R. Waitman, and M. Liu, "Feature ranking in predictive models for hospital-acquired acute kidney injury," *Scientific Reports*, vol. 8, Nov. 2018.

[10] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "Interpretability in healthcare: A comparative study of local machine learning interpretability techniques," *Computational Intelligence*, vol. 37, pp. 1633–1650, Nov. 2020.

[11] R. Elshawi, M. H. Al-Mallah, and S. Sakr, "On the interpretability of machine learning-based model for predicting hypertension," *BMC Medical Informatics and Decision Making*, vol. 19, July 2019.

[12] I. Ilic, B. Görgülü, M. Cevik, and M. G. Baydoğan, "Explainable boosted linear regression for time series forecasting," *Pattern Recognition*, vol. 120, p. 108144, 2021.

[13] D. Barić, P. Fumić, D. Horvatić, and T. Lipic, "Benchmarking attention-based interpretability of deep learning in multivariate time series predictions," *Entropy*, vol. 23, p. 143, Jan. 2021.

[14] D. A. Kaji, J. R. Zech, J. S. Kim, S. K. Cho, N. S. Dangayach, A. B. Costa, and E. K. Oermann, "An attention based deep learning model of clinical events in the intensive care unit," *PLOS ONE*, vol. 14, p. e0211057, Feb. 2019.

[15] C. Chen and B. Li, "An interpretable channelwise attention mechanism based on asymmetric and skewed gaussian distribution," *Pattern Recognition*, p. 109467, 2023.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014.

[18] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, *et al.*, "Explainable ai (xai): Core ideas, techniques, and solutions," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[19] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[20] J. Zacharias, M. von Zahn, J. Chen, and O. Hinz, "Designing a feature selection method based on explainable artificial intelligence," *Electronic Markets*, pp. 1–26, 2022.

[21] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep ppg: large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, no. 14, p. 3079, 2019.

[22] O. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, and C. Villalonga, "mHealthDroid: A novel framework for agile development of mobile health applications," in *Ambient Assisted Living and Daily Activities*, pp. 91–98, Springer International Publishing, 2014.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] J. Ni, L. Muhlstein, and J. McAuley, "Modeling heart rate and activity data for personalized fitness recommendation," in *The World Wide Web Conference*, pp. 1343–1353, 2019.

[25] Z. Zhu, H. Li, J. Xiao, W. Xu, and M.-C. Huang, "A fitness training optimization system based on heart rate prediction under different activities," *Methods*, vol. 205, pp. 89–96, 2022.

[26] T. Shaik, X. Tao, N. Higgins, R. Gururajan, Y. Li, X. Zhou, and U. R. Acharya, "FedStack: Personalized activity monitoring using stacked federated learning," *Knowledge-Based Systems*, vol. 257, p. 109929, Dec. 2022.

[27] F. Bozkurt, "A comparative study on classifying human activities using classical machine and deep learning methods," *Arabian Journal for Science and Engineering*, vol. 47, pp. 1507–1521, July 2021.

[28] L. Yijing, Y. Wenyu, Y. Kang, Z. Shengyu, H. Xianliang, J. Xingliang, W. Cheng, S. Zehui, and L. Mengxing, "Prediction of cardiac arrest in critically ill patients based on bedside vital signs monitoring," *Computer Methods and Programs in Biomedicine*, vol. 214, p. 106568, 2022.

[29] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, p. 104672, Sept. 2021.

[30] S. Malakar, S. D. Roy, S. Das, S. Sen, J. D. Velásquez, and R. Sarkar, "Computer based diagnosis of some chronic diseases: A medical journey of the last two decades," *Archives of Computational Methods in Engineering*, pp. 1–43, 2022.

[31] L.-V. Herm, K. Heinrich, J. Wanner, and C. Janiesch, "Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability," *International Journal of Information Management*, vol. 69, p. 102538, 2023.