A Deep Neuro-Fuzzy Method for ECG Big Data Analysis via Exploring Multimodal Feature Fusion

Xiaohong Lyu, *Member, IEEE,* Shalli Rani, *Senior Member, IEEE,* S. Manimurugan, *Senior Member, IEEE,* Carsten Maple, *Senior Member, IEEE*, Yanhong Feng, *Member, IEEE*

***Abstract*—In the realm of medical data processing, particularly in the diagnosis and monitoring of cardiac diseases, the analysis of Electrocardiogram (ECG) signals represents a critical challenge, especially with the burgeoning volume of ECG big data. Traditional methods and existing research often fall short in effectively analyzing this data, limited by their inability to fully capture the complex and nonlinear patterns inherent in ECG signals. Addressing these limitations, in this paper, we introduce a novel deep neuro-fuzzy model augmented with multimodal feature fusion. Our method ingeniously combines the power of neuro-fuzzy systems with the robust feature extraction capabilities of deep learning, specifically leveraging a Transformer-based architecture, to analyze both ECG signals and their corresponding spectral images. This multimodal fusion not only enriches the model's input data, providing a comprehensive understanding of cardiac signals, but also enhances the adaptability and accuracy of cardiac arrhythmia detection. We rigorously validate our approach on the MIT-BIH Arrhythmia Database, conducting a series of experiments, including performance evaluations and ablation studies, to highlight the significant contributions of the multimodal feature fusion and neuro-fuzzy module. The results achieve significant improvements in classification metrics: an accuracy of 98.46% and an F1 score of 99.1%. Moreover, we benchmark the Transformer's feature extraction performance against other architectures like ResNet. The results unequivocally demonstrate our model's superiority and illustrate the potential of integrated neuro-fuzzy and deep learning approaches in overcoming the current limitations of ECG signal analysis.**

***Index Terms*—Medical data analysis, deep neuro-fuzzy network, multimodal feature, ECG analysis.**

I. Introduction

In the past few decades, with the rapid development of biomedical engineering and information technology, electrocardiogram (ECG) has become one of the most commonly used and non-invasive technologies in the diagnosis of heart diseases [1]. The large-scale collection and analysis of electrocardiogram data, commonly known as ECG big data, provides unprecedented opportunities for early diagnosis and treatment of heart diseases. However, the high dimensionality and complexity of ECG big data pose great challenges to data processing and analysis.

Traditional ECG data analysis methods mainly focus on one-dimensional signal processing. Although these methods are effective, they often encounter bottlenecks when processing large-scale data sets. In recent years, the development of deep learning technology has provided new perspectives and methods for the analysis of ECG data. In particular, great potential has been shown by converting ECG signals into spectral images and analyzing these data using image processing techniques. In addition, multi-modal feature fusion technology, which comprehensively utilizes the features of one-dimensional data signals and two-dimensional image data, is considered an effective means to improve the accuracy of ECG data analysis.

Deep neuro-fuzzy models combine the powerful feature extraction capabilities of deep learning with the uncertainty handling ability of neuro-fuzzy systems [2], [3]. Neuro-fuzzy systems, intelligent systems that mimic human decision-making processes using fuzzy logic principles, have been widely applied in modeling and controlling various complex systems due to their ability to handle uncertainty and imprecision [4]. When combined with deep learning technologies, they not only learn deep, abstract features from vast amounts of data but also process information's uncertainty through fuzzy rule systems, thereby enhancing model interpretability and adaptability. By integrating the strengths of deep learning models with neuro-fuzzy systems, the proposed deep neuro-fuzzy method effectively enhances the performance of ECG big data analysis. This approach can process and analyze multimodal features extracted from ECG signals, including one-dimensional signals and two-dimensional spectral images. It leverages the deep learning model's powerful feature extraction capabilities and the neuro-fuzzy system's advantage in handling uncertainties. This method not only strengthens the model's ability to recognize complex patterns in ECG signals but also improves the transparency and interpretability of the decision-making process, which is particularly important for applications in the healthcare sector.

Multimodal data fusion refers to the process of integrating information from multiple data sources or modalities to improve the understanding, learning, or decision-making process in a computational model [5]. This approach leverages the complementary strengths of different data types, such as text, images, audio, or sensor data, to achieve a more comprehensive and nuanced analysis than what could be obtained

This work was supported by the Natural Science Foundation of Liaoning Province (Doctoral Research Initiation) under Granted NO. 2023-BSBA-141. Corresponding author: Yanhong Feng.

X. Lyu and Y. Feng are with the First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121012, China (e-mail: lvxh@jzmu.edu.cn, fengyh@jzmu.edu.cn).

S. Rani is with the Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab-140401, India (e-mail: shalli.rani@chitkara.edu.in).

S. Manimurugan is with the Faculty of Computers and Information Technology, University of Tabuk, Tabuk 47512, Saudi Arabia (e-mail: mmurugan@ut.edu.sa).

C. Maple is with the Secure Cyber Systems Research Group, WMG, University of Warwick, Coventry CV74AL, U.K. (e-mail: cm@warwick.ac.uk).

from any single data type alone. In the context of ECG analysis, combining one-dimensional (1D) ECG signals with two-dimensional (2D) spectral images is a powerful example of multimodal data fusion. ECG signals, which measure the electrical activity of the heart, provide valuable information about cardiac health. However, transforming these signals into spectral images can unveil additional features not readily apparent in the time domain, such as frequency components and energy distribution over time. This multimodal approach enhances the model's ability to diagnose and analyze cardiac health issues, offering a more nuanced understanding than could be achieved by analyzing the ECG data or spectral images alone. For instance, a model might use the 1D signal to identify abnormal heart rhythms while simultaneously using the spectral image to investigate the presence of specific arrhythmias or other heart conditions that manifest in the frequency domain.

In recent years, the prevalence of cardiac diseases has surged, making timely and accurate diagnosis crucial for effective treatment and patient management. Electrocardiogram (ECG) signals, as a primary tool for monitoring heart activity, provide invaluable insights into cardiac health. However, traditional methods for ECG analysis often struggle to fully capture and interpret the complex, nonlinear patterns inherent in these signals, particularly when dealing with large-scale data. Existing algorithms, while effective to a certain extent, are often limited by their reliance on single-modal data and lack the capacity to handle the intricate variability and uncertainty present in ECG signals. Motivated by these challenges, our research aims to address these limitations through the innovative integration of deep learning and neuro-fuzzy systems, augmented by multimodal feature fusion. By combining the strengths of Transformer-based architectures and the interpretability of fuzzy logic, we propose a comprehensive approach that enhances the accuracy and robustness of ECG signal classification. This work not only advances the field of biomedical signal processing but also has significant implications for improving the diagnosis and monitoring of cardiac conditions, ultimately contributing to better patient outcomes.

In this paper, we propose a deep neuro-fuzzy method for multi-modal feature fusion analysis of ECG big data. First, we convert the ECG signal into a spectrum image and use the converted image to capture the time and frequency characteristics of the signal. Next, we designed a feature extraction encoder based on the Transformer model to extract features from one-dimensional ECG data signals and spectrum images. The encoder is able to effectively process and fuse data from these two modalities, capturing the deep features of the signal. Finally, we use a neuro-fuzzy model to analyze the fused multi-modal features to achieve high-accuracy classification and prediction of ECG data. The combination of multimodal feature fusion, advanced Transformer-based feature extraction, and neuro-fuzzy system integration sets our proposed algorithm apart from existing methods. This holistic approach not only enhances the accuracy and robustness of ECG signal classification but also provides a framework that is adaptable, addressing key challenges in the analysis of ECG big data. This research not only provides a new perspective and methodology for the field of ECG big data analysis but also explores the beneficial integration of deep learning and neuro-fuzzy systems. Through in-depth study and experimental validation, we demonstrate the effectiveness and superiority of the proposed method in ECG data analysis, significantly contributing to the advancement of healthcare data analysis.

The main contributions of this paper include as follows:

- We propose a novel ECG analysis model incorporating neuro-fuzzy systems. This model harnesses the interpretability of fuzzy logic and the learning capacity of neural networks, offering an adaptive solution to the complexities of ECG data, thereby enhancing cardiac abnormality detection and classification.
- We introduce a sophisticated method for the extraction and fusion of multimodal features from both one-dimensional ECG signals and two-dimensional spectral images. Our approach capitalizes on the distinct yet complementary information provided by each modality, leading to a more comprehensive analysis of cardiac activity and improving the detection of subtle disease indicators.
- We rigorously validate our model and feature fusion method through extensive experiments. Our testing not only confirms the model's accuracy in classifying cardiac conditions but also showcases its robustness across diverse patient datasets. Our findings demonstrate the superiority of our approach over traditional and current deep learning techniques in ECG data analysis.

The organization of the paper is as follows. Section 2 details related work, including traditional methods and deep learning methods for ECG data analysis, as well as the latest progress in multi-modal feature fusion technology. Section 3 describes the specific implementation of the method proposed in this paper, including data preprocessing, design of feature extraction encoder, and construction of neuro-fuzzy model. Section 4 demonstrates the performance of the proposed method through experimental results and compares it with other existing techniques. Finally, Section 5 summarizes the full paper.

II. Related Work

In the related work section of our study, we embark on a comprehensive review of the existing literature, focusing on the evolution of ECG data analysis techniques and the burgeoning field of multimodal data fusion. This exploration is essential to contextualize our research within the broader academic discourse, highlighting how our contributions address gaps and introduce novel approaches to the challenges inherent in cardiac data analysis.

Historically, ECG data analysis has leveraged a range of methodologies, from traditional signal processing techniques to advanced machine learning and deep learning models. These approaches have laid a solid foundation for interpreting the intricate patterns of ECG signals, facilitating the diagnosis of various cardiac conditions. However, as the complexity of the data and the demand for precision in diagnostics has increased, researchers have turned to more sophisticated models that can capture the nuanced relationships within cardiac signals and between different types of data modalities.

In traditional research methods for classifying ECG signals, the majority of studies begin with the extraction of features across the time domain, frequency domain, and time-frequency domain [6], [7]. Following this foundational step, some research extends into the extraction of nonlinear features [8], [9]. Time domain analysis focuses on the ECG signal's fundamental shape and rhythm, highlighting elements such as the R wave peak, RR interval, P wave, and T wave. Frequency domain analysis elucidates the signal's frequency components, encompassing aspects like ECG spectral density, power spectrum, and waveform decomposition. Time-frequency domain analysis combines insights from both time and frequency domains, offering a more comprehensive view of ECG signal features. This approach includes techniques such as the short-time Fourier transform [10]–[12] and wavelet transform [13], [14]. Within machine learning applications, these features play a crucial role in the extraction, classification, and analysis of ECG signals, aiding in the identification and categorization of cardiac events to support clinical diagnosis and treatment.

Ref. [15] described a method using machine learning technology for arrhythmia detection. This method used discrete wavelet transform [16] to process the ECG signal, extracted features from it, and used these features to identify five types of arrhythmic beats, and then combined independent component analysis technology to perform feature dimensionality reduction. Finally, a support vector machine classifier is used for classification. After experiments on the MIT-BIH arrhythmia database, an average accuracy rate of over 98% was obtained. Ref. [17] used sparse representation technology to decompose the ECG signal into basic waves and extract four features such as time delay, frequency, width, and expansion coefficient square, and then splice and analyze these features. Finally, the least squares twin support vector machine with particle swarm optimization is used to further classify these features to achieve efficient analysis of ECG signals. The above method was tested using the MIT-BIH arrhythmia database and achieved an accuracy of over 99%.

Deep learning algorithms are highly penetrated in the field of cardiovascular disease classification and prediction. Due to the variable data length of ECG signals, Convolutional Neural Networks (CNN) have become in some form the preferred architecture for this type of task. Attia *et al.* [18] used standard 12-lead electrocardiogram to detect atrial fibrillation through convolutional neural network. Zhu *et al.* developed a CNN model to generate diagnoses that can distinguish 21 arrhythmias based on single-label and multi-label electrocardiograms [19]. Xie *et al.* proposed a convolutional neural network based on discrete biorthogonal wavelet transform for ECG atrial fibrillation detection [20]. Xiong *et al.* proposed a method for detecting inferior wall myocardial infarction based on densely connected convolutional neural networks [21]. Al-Zaiti *et al.* [22] used the spatiotemporal features of standard 12-lead electrocardiograms to predict acute myocardial ischemia in patients with chest pain through supervised learning through different machine learning classifiers. Mostayed *et al.* used a bidirectional LSTM network to classify and predict arrhythmias on 12-lead ECG signals in the Physiological Signal Challenge [23]. The proposed model can be trained on ECG signals of any length. Yao *et al.* integrated CNN, LSTM and attention modules to achieve spatial and temporal fusion of ECG signal information to detect paroxysmal arrhythmias [24]. Saadatnejad *et al.* adopted a novel architecture composed of wavelet transform and multiple LSTMs to automatically extract features and improve heartbeat classification accuracy [25]. The framework has low computational costs and meets time requirements for continuous execution on wearable devices with limited processing capabilities. Wang *et al.* [26] introduced an interpretable arrhythmia classification method from a new perspective, called human-computer collaborative knowledge representation, which can not only give classification results but also provide the characteristic basis for classification. Xu *et al.* proposed a combined network of convolutional neural networks and recurrent neural networks, and the developed network also achieved high accuracy when detecting 95 electrocardiogram categories [27].

The CNN in deep learning models can automatically and implicitly learn features from training data, which improves the objectivity of the extracted features and eliminates the subjective biases associated with manual feature extraction [28]–[30]. According to Ref. [31], three neural network architectures were proposed: one based on convolutional networks, another on SincNet, and a third combining convolutional networks with entropy features. These architectures classified 2, 5, and 20 categories of diseases using the PTB-XL dataset, respectively, with the convolutional network incorporating entropy features achieving the best classification results in all tasks. Ref. [32] utilized a method that integrates autoencoders with self-organizing maps, achieving high accuracy in experiments on the MIT-BIH arrhythmia database and the PTB ECG database. Ref. [33] proposed a new model of wavelet sequence based on a deep bidirectional long short-term memory network for classifying ECG signals. Ref. [34] proposed an attention-based time incremental convolutional neural network, which realizes the spatial and temporal fusion of ECG signal information by integrating the convolutional neural network, recurrent unit, and attention module.

In recent years, the integration of multimodal data sources, such as combining one-dimensional ECG signals with two-dimensional spectral images, has emerged as a promising avenue for enhancing the accuracy and robustness of ECG analysis. This shift towards multimodal fusion acknowledges the complementary nature of different data types and seeks to leverage this synergy to achieve a more holistic understanding of cardiac health. Furthermore, the application of neuro-fuzzy systems in the realm of ECG analysis represents an innovative step forward, marrying the human-like reasoning capabilities of fuzzy logic with the adaptability and learning prowess of neural networks [35]. This hybrid approach aims to address some of the limitations of purely data-driven models, such as their often opaque decision-making processes and challenges in handling noisy or incomplete data. In recent years, the integration of deep learning with neuro-fuzzy systems has garnered significant attention due to its potential to enhance the interpretability and robustness of complex models. Several studies have explored the application of deep neuro-fuzzy models in various domains, including ECG data processing. Our proposed approach advances the state-of-the-art by integrating multimodal feature fusion. Unlike previous models, our method captures both time-domain and frequency-domain information from ECG signals. Additionally, the neuro-fuzzy system in our model enhances robustness, addressing the limitations of existing approaches.

Disease detection based on single-modal medical data is still the most conventional method at present, but there are also shortcomings of a single type of information. Medical diagnostic research based on machine learning is developing in a multi-modal direction. Currently, researchers are extensively exploring potential ways to combine multiple modalities to detect cardiovascular disease. For example, Ref. [36] proposed two experiments to systematically investigate the impact of ECG duration and multimodal information on model training, and introduced a novel multimodal deep learning architecture to learn joint features from both ECG and patient demographics. Sharma et al. [37] introduced a multi-modal neural network component that utilizes both MRI and clinical data to differentiate between healthy individuals and those with myocardial infarction. Schwob et al. [38] developed several hybrid neural networks designed to analyze various modalities, achieving reliable results in arrhythmia detection. Zarrabis et al. [39] investigated an automatic diagnosis method for acute myocardial infarction using multi-modal data, collecting ECG signals, heart sound signals, and clinical characteristics to design an automatic diagnosis model. Xiao et al. [40] proposed a novel feature fusion deep learning architecture for the early identification of myocardial infarction, leveraging electrocardiogram and patient demographic data. Ahmad et al. [41] employed Gram angle fields, recurrence maps, and Markov transfer fields to convert raw ECG data into three distinct image types, using a multi-modal image fusion framework for arrhythmia and myocardial infarction classification. Phan et al. [42] proposed a multimodal ECG arrhythmia classification model based on self-supervised learning (SSL). They utilized self-knowledge distillation technology to train on the time and frequency domains of unlabeled data in the SSL pre-training task and introduced a gate fusion mechanism to merge information from multiple modalities.

III. Method

In the methodology section of our paper, we meticulously outline the steps and processes involved in our innovative approach to ECG data analysis. First, we detail the method for converting ECG signals into spectrograms. Our approach integrates multimodal feature fusion, specifically combining one-dimensional ECG signals and their corresponding two-dimensional spectral images. This fusion allows us to capture both time-domain and frequency-domain information, which provides a more comprehensive representation of the cardiac signals. Second, we introduce a feature extraction encoder based on the Transformer model for extracting features from one-dimensional ECG data signals and spectrum images. While Transformers have been widely used in natural language processing and other domains, their application to ECG signal analysis, particularly in combination with fuzzy logic, is novel. The use of Transformers allows us to effectively model long-range dependencies and complex patterns within the ECG data. Third, we use a neuro-fuzzy model to analyze the fused multi-modal features to achieve high-precision classification and prediction of ECG data. The integration of the neuro-fuzzy system enhances the model's predictions, addressing a critical challenge in the application of deep learning to medical data. The neuro-fuzzy module provides a mechanism to handle the inherent uncertainties and variabilities in ECG signals, improving the robustness of the diagnostic process. The overall frame diagram of the model is shown in Fig. 1.

A. Image Transformation

This section is dedicated to describing how we transform ECG signals into spectral images using the Short-Time Fourier Transform (STFT). This transformation is a pivotal initial step in our multimodal data analysis strategy, enabling the fusion of one-dimensional ECG signals with two-dimensional spectral representations for enhanced feature extraction and analysis. The STFT is employed to analyze the frequency components of the ECG signals over time, providing a dynamic view of how these components change. This is particularly important for capturing the transient behaviors in ECG data that traditional Fourier Transform might miss due to its assumption of signal stationarity. Considering that the ECG signal is an unbalanced signal, Fourier transform cannot be used directly to obtain frequency domain features. Therefore, we use the STFT to transform a heartbeat signal from the time domain to the frequency domain. By segmenting the signal into shorter frames and applying Fourier Transform to each, the STFT generates a time-frequency representation of the ECG signal, which is then visualized as a spectral image. This spectral image not only encapsulates the original signal's temporal and frequency information but also unveils patterns and anomalies that may not be visible in the raw ECG data. The transition from one-dimensional signal analysis to a two-dimensional spectral analysis opens up new avenues for feature extraction, leveraging both the spatial relationships within the images and the temporal dynamics of the ECG signals.

The main idea of STFT transform is to first decompose the signal into countless small segments, and then add window processing to each small segment. The windowed small segment signal can be considered as a stationary signal, and then use Fourier transform to process the signal. Suppose the signal is $x(t)$, then the STFT transformation of the signal is:

$$STFT_x(t,f) = \int x(\tau)h(\tau - t)\mathrm{e}^{-\mathrm{j}2\pi f\tau}\mathrm{d}\tau, \qquad (1)$$

where t is time, f is frequency, and $h(t)$ is the window function. In order to perform Fourier transform processing on the signal on the local time-frequency plane, a limited time window function $h(t)$ is added before the Fourier transform, so the window function $h(t)$ divides the non-stationary signal into countless time segments. Under the action of the window

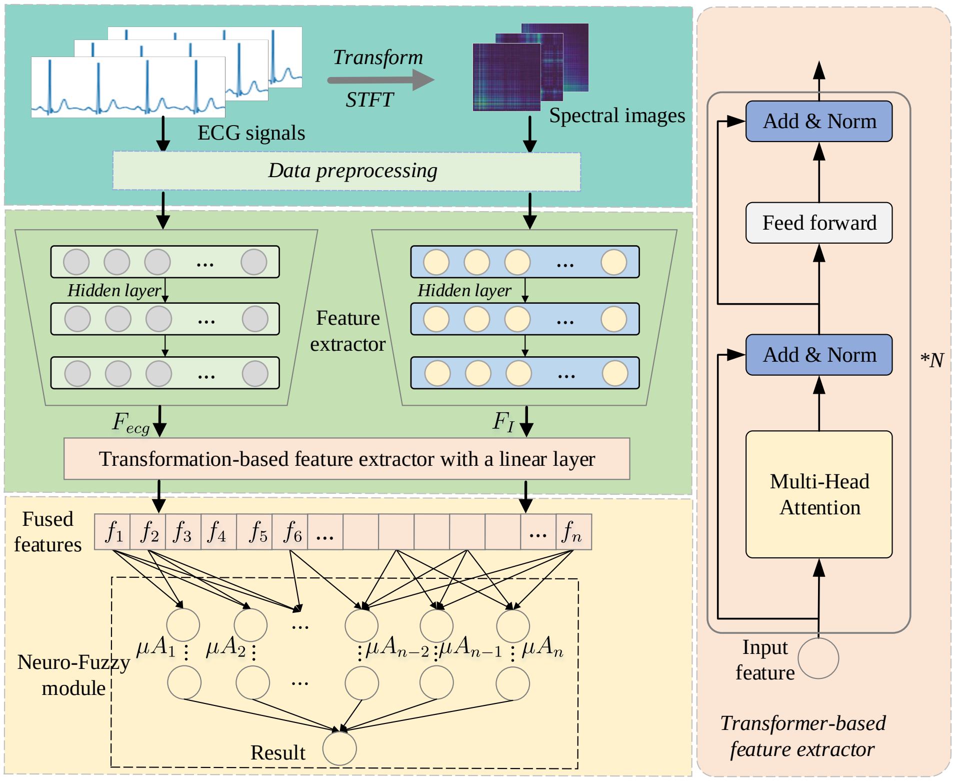

Fig. 1. The overall structure of the proposed ECG analysis model incorporating neuro-fuzzy system.

function $h(t)$, we perform Fourier transformation on each small segment of the signal, thereby obtaining the ”local” spectrum of the ECG signal. The time-varying characteristics of the signal can be obtained from the ”local” spectrum differences at different times.

The frequency of each small segment of signal is determined under the action of window function $h(t)$. The smaller the window of the window function, the higher the resolution. Since the time width and bandwidth of the window are contradictory and also limit the size of the window, the window cannot be arbitrarily small. However, after windowing the signal $x(t)$, the result of the Fourier transform can be connected with the time variable. The window function of the STFT transform is calculated as follows.

$$h^{(p,q)}(t) = \mathrm{e}^{jpt} h(t-q), h(t) \in L^2(R), \tag{2}$$

where p represents the frequency shift parameter of the window, and q represents the time shift parameter of the window. In practical operations, we take $p = mp_0$ and $q = nq_0$. At this time, the STFT transformation of the signal is:

$$C_{m,n}(x) = \{h_{m,n}, x\} = \{h_{mp_0,nq_0}, x\} \tag{3}$$

where $C(x)$ is the coefficient of the STFT transformation. As can be seen from the above equation, when $mp_0, nq_0 (m, n \in Z)$ changes, the signal is cut into a series of equal rectangular blocks at equal intervals by the variable h_{mp_0,nq_0}. The equal resolution of the signal mapped by $L^2(R) - L^2(R^2)$ to the time-frequency plane is an important feature of the STFT transform.

The advantage of the STFT transform is that it is easy to characterize the signal, and the piecewise stationarity is the characteristic of the analyzed signal. In order to obtain the best analysis results for stationary segments of different time scales, we need to set the window length and type that are compatible with the stationary segment signals. The frequency domain solution of STFT transform is inversely proportional to the time domain solution. Improves the frequency domain solution, but makes the time domain solution poor. Improves the time domain solution, but makes the frequency domain solution worse. Therefore, it is impossible to obtain frequency domain solutions and time domain solutions at the same time by using STFT transform. This is a shortcoming of STFT transform in the application process.

After performing STFT on a heart beat signal, the spectrum we get is a $d_t \times d_f$ dimensional matrix, which is a function of time (corresponding to different window movements) and frequency. Each column in it is the frequency distribution of the signal within the current window function time. The column vectors in the spectrum data are arranged in time order to form a matrix, and each vector represents the signal characteristics in the corresponding time window. The dependence relationships in the two dimensions of time and frequency in the spectrum matrix corresponding to different heart beat types are different, thus expressing richer data information.

By integrating the STFT into our methodology, we lay the foundation for a comprehensive multimodal analysis, setting the stage for the feature extraction and fusion processes that follow. This subsection not only explains the rationale behind using the STFT but also highlights its significance in enhancing the depth and breadth of our ECG data analysis approach.

B. Feature Extraction based on Transformer

In this section, we delve into the design of our feature extraction module. We utilize Transformer architectures as the backbone to extract salient features from both the ECG signals and their corresponding spectral images. This dual-modal approach leverages the Transformer's ability to capture long-range dependencies and complex patterns within the data, making it particularly suited for analyzing the intricate structures of ECG signals and spectral images. The Transformer model, initially introduced for natural language processing tasks, has been widely adopted for its effectiveness in handling sequential data. Transformer [43] is suitable for end-to-end tasks such as sequence to sequence (seq2seq). Transformer-based models not only have good performance on natural language processing problems, but also perform well on computer vision classification, detection, and segmentation tasks [44]. The traditional Transformer contains a set of Encoders and a set of Decoders, as shown in Fig. 1. The most important structure in Encoder and Decoder is Self-Attention. Through the Self-Attention mechanism, Transformer can capture global dependencies. This is also one of the reasons why Transformer is better than CNN and RNN in processing sequence problems. Its core mechanism, the self-attention mechanism, allows the model to weigh the importance of different parts of the input data, enabling it to focus on relevant features while minimizing the impact of less informative ones.

For our application, we adapt the Transformer architecture to analyze both one-dimensional ECG signals and two-dimensional spectral images. The model consists of two parallel branches, each optimized for its respective data modality. Next, we detail the computation process of data within our Transformer-based feature extraction model for both ECG signals and their corresponding spectral images, leading to the integration of features through fusion. This process involves a series of computational steps designed to extract and merge the most informative features from both modalities for further analysis.

For the ECG signal analysis branch, the Transformer processes the input signal S composed of N data points, resulting in a sequence of feature vectors $x = [x_1, x_2, ..., x_N]$. Each x_i is a feature representation of a time step. The computation in the Transformer layer for the ECG branch is primarily driven by the self-attention mechanism, which is defined as follows. The self-attention mechanism computes the attention scores between different positions of the input sequence to capture the interdependencies regardless of their positions as follows:

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where Q, K, and V are the query, key, and value matrices derived from the input, and d_k is the dimension of the key, and their calculations is:

$$Q = W_q \times p, \tag{5}$$

$$K = W_k \times p, \tag{6}$$

$$V = W_v \times p, \tag{7}$$

where W_q, W_k, and W_v are weight matrices randomly initialized at the beginning of network training, their value can be adjusted through backpropagation of the network, and p is the input to the self-attention layer.

For multi-head attention, this process is performed in parallel with different, learned linear projections of the inputs, enhancing the model's ability to focus on different aspects of the input sequence. Multi-Head Attention is an extension of the attention mechanism, which uses multiple parallel attention heads to improve the model's ability to extract data features. Each head can capture a different focus of attention, analogous to how different people pay different attention to a picture. At the same time, the multi-head attention mechanism and multi-kernel convolution have certain similarities, and they capture different features respectively. The specific implementation is the same as the single-head attention mechanism. It is necessary to apply for several more sets of weight matrices. The calculation is as follows:

$$head_i = \text{Att}\left(QW_q^i, KW_k^i, VW_v^i\right) \tag{8}$$

$$\text{MultiHead}\left(Q, K, V\right) = \text{Concat}\left(head_1, ..., head_\text{h}\right)W^o \tag{9}$$

After attention calculation, the output for each position passes through a feed-forward network applied independently to each position, i.e., the position-wise feed-forward networks:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{10}$$

This is applied to each element in the sequence, ensuring that the model captures the local context around each point in the signal.

For the spectral image analysis branch, the input spectral image I is represented as a sequence of vectors $y = [y_1, y_2, ..., y_M]$, where each y_j corresponds to a flattened patch or a processed region of the image. The Transformer processes these vectors similarly to the ECG signal analysis, employing self-attention and feed-forward networks to capture spatial and frequency-domain features. The calculation follows the same formulas as above but is applied to the image data representation.

C. Multimodal Fusion and Fuzzy Module

In this section, we focus on the process of fusing the features extracted from both ECG signals and spectral images and subsequently analyzing these features using a neuro-fuzzy network to derive the final diagnostic results. This approach leverages the complementary nature of the two modalities and employs the neuro-fuzzy network's capability to handle the imprecision and uncertainty inherent in medical data, thereby improving the robustness and accuracy of cardiac condition diagnosis.

Upon obtaining the feature representations F_{ECG} and $F_{Spectral}$, the next step is to effectively merge these features into a unified representation that captures the holistic insights from both the time-domain and frequency-domain data. The fusion process can be formalized as follows:

$$F_{fused} = \phi(F_{ecg} \oplus F_I) \tag{11}$$

where $\oplus$ denotes the fusion operation (e.g., concatenation, element-wise addition), ϕ represents a learnable transformation (such as a linear layer or a neural network module) designed to integrate and compress the information from the concatenated features into a dense representation F_{fused}.

The neuro-fuzzy network, integrating the principles of fuzzy logic with neural network learning, is utilized to analyze the fused features F_{fused} for the final diagnostic output. The core idea is to model the fuzzy inference system within a neural network framework, allowing the system to learn the fuzzy rules and membership functions from the data. The general architecture of the neuro-fuzzy network includes the fuzzification layer and the rule layer. The fuzzification layer converts the input features F_{fused} into fuzzy values by applying membership functions. Each input feature is associated with several fuzzy sets $\mu_{A_i}(f)$, with each set represented by a membership function indicating the degree to which the input belongs to the set. $\mu_{A_i}(f)$ is the membership function of fuzzy set A_i for input f. The rule layer implements the fuzzy logic rules. Each neuron in this layer represents a fuzzy rule. The inputs to the rule layer are the fuzzy values from the fuzzification layer, and the output is the rule strength, typically calculated using the AND (min) or OR (max) operator for the antecedents of the rule, as follows.

$$R_j = \min(\mu_{A_1}(f_1), \mu_{A_2}(f_2), ..., \mu_{A_n}(f_n)) \tag{12}$$

where R_j is a rule with antecedents $A_1, A_2, ..., A_n$.

The defuzzification layer converts the fuzzy rule strengths back into a crisp output value, providing the final diagnostic prediction. This is often achieved through a weighted average or a centroid method, where the outputs of the rule layer are combined based on their contribution to the output fuzzy sets, as follows.

$$\hat{y}_i = \frac{\sum_{j=1}^{m} R_j \cdot y_j}{\sum_{j=1}^{m} R_j} \tag{13}$$

where $\hat{y}_i$ is the output, R_j is the strength of rule j, and y_j is the output value associated with rule j.

Through this neuro-fuzzy model, the system can accommodate the nuances and uncertainties of the ECG and spectral image data, leading to a more accurate and interpretable diagnostic outcome. The learnable parameters within the neuro-fuzzy network, including the membership functions and rule weights, are optimized during training to minimize the discrepancy between the network's predictions and the actual diagnoses, using a dataset of annotated ECG recordings and images. This optimization allows the model to adaptively improve its diagnostic capability, making it a powerful tool for cardiac health assessment.

In order to understand the model details more clearly, we explain the overall process of the framework as follows. Our proposed model integrates a Transformer-based deep learning architecture with a neuro-fuzzy system to enhance the analysis of ECG signals. The process begins with the extraction of features from both the ECG signals and their corresponding spectral images using the Transformer. For feature extraction, the raw ECG signals are normalized and fed into a Transformer model, which processes the sequential data to extract high-level temporal features. The ECG signals are also transformed into spectral images using the STFT. These images are then processed by another Transformer model to extract frequency-domain features. For multimodal feature fusion, the extracted features from both the ECG signals and the spectral images are concatenated to form a comprehensive feature set. This concatenation allows the model to leverage the complementary information from both modalities, providing a richer representation of the cardiac signals. For the neuro-fuzzy system integration, the fused features are fed into a fuzzification layer, where they are converted into fuzzy values using predefined membership functions. These functions define how each feature contributes to various fuzzy sets. The fuzzy values are then processed by a set of fuzzy rules in the inference engine. Each rule evaluates the degree to which the input features satisfy certain conditions, producing a corresponding output value. The output values from the fuzzy inference engine are aggregated and defuzzified to produce a crisp output, which represents the predicted class of the ECG signal. The multimodal feature fusion is implemented by concatenating the feature vectors obtained from the Transformer models processing the ECG signals and the spectral images. This fusion step is crucial as it allows the model to integrate both temporal and frequency-domain information, leading to a more holistic understanding of the cardiac signals. The overall pseudocode flow of our method is shown in Algorithm 1.

Algorithm 1 Deep Neuro-Fuzzy Method with Multimodal Feature Fusion for ECG Signal Classification

```
Input: ECG signal dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ is
    the $i$-th ECG signal and $y_i$ is its corresponding label;
Output: Predicted labels $\hat{y}_i$ for each ECG signal;
for each signal $\mathbf{x}_i$ in dataset $\mathcal{D}$ do
    Normalize ECG signals $\mathbf{x}_i$ using min-max normalization;
    Apply STFT to obtain spectral images;
end for
for each normalized $\mathbf{x}_{i,norm}$ in dataset $\mathcal{D}$ do
    Apply Transformer-based model to extract features from $\mathbf{x}_{i,norm}$;
    Apply Transformer-based model to extract features from spectral images;
end for
Get extracted features $F_{ecg_i}$, $F_{I_i}$;
for each set of features $F_{ecg_i}$ and $F_{I_i}$ do
    Fuse features using concatenation according to Eq. (11);
    Get fused features $F_{fused}$;
    Apply fuzzification to $F_{fused}$ to obtain fuzzy values according to Eq. (12);
    Apply defuzzification to obtain $\hat{y}_i$ according to Eq. (13);
end for
for each crisp output $y_i$ do
    Predict the class of the ECG signal;
end for
return Predicted labels $\{\hat{y}_i\}_{i=1}^{N}$
```

IV. Experiments

In this section, we meticulously outline our experimental setup, conduct a comprehensive performance comparison with existing methodologies, and delve into ablation studies to validate the significance of the design choices in our proposed model. This systematic approach allows us to demonstrate the efficacy and robustness of our neuro-fuzzy model integrated with multimodal feature fusion for ECG data analysis.

A. Experimental Configurations

Initially, we describe the experimental environment, including the dataset specifications, preprocessing steps, and the parameter settings of our model. We utilize a diverse collection of ECG datasets, ensuring a broad representation of cardiac conditions to challenge our model's diagnostic capabilities. The preprocessing phase involves signal normalization, noise filtering, and the conversion of ECG signals into spectral images, setting the stage for feature extraction. Additionally, we detail the configuration of our Transformer-based feature extraction module and neuro-fuzzy network, including the architecture details, learning rates, and other hyperparameters crucial for replicating our experiments.

1) Environment: In our research, the experimental setup is designed to support the rigorous demands of deep learning experiments. Our hardware foundation consists of an NVIDIA GeForce RTX 3080 GPU, alongside an Intel Core i9-10900K CPU and 32 GB of DDR4 RAM, ensuring smooth data processing and adequate memory for handling large datasets. Storage is facilitated through a 1 TB NVMe SSD, speeding up data retrieval and storage operations. Software-wise, we operate within Ubuntu 20.04 LTS. Python 3.8 serves as our programming base. PyTorch 1.8.1, our chosen deep learning framework, enables dynamic modeling and experimentation, supported by CUDA 11.1 and cuDNN 8.0.4 for GPU acceleration. Environment and dependency management is streamlined using Conda, whereas Visual Studio Code, enhanced with the Python extension, offers a conducive environment for code development and debugging. This configuration ensures a balanced and efficient platform for conducting our deep learning research, offering both the computational power and the software flexibility required to push the boundaries of current methodologies.

In our study, we carefully tuned the hyperparameters of our proposed network model to optimize performance on the ECG signal classification task. The Transformer architecture used in our model consisted of 4 encoder layers, each with 8 attention heads, an embedding dimension of 128, and a feedforward network dimension of 512. We applied a dropout rate of 0.1 to prevent overfitting. The learning rate was set to 0.001, optimized with a learning rate scheduler. Training was conducted with a batch size of 64 over 100 epochs, employing early stopping based on validation performance. The Adam optimizer was utilized with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$.

2) Database and Metrics: The MIT-BIH Arrhythmia Database serves as the cornerstone of our experimental dataset, a choice motivated by its extensive adoption and validation within the biomedical signal processing community. Established by the Massachusetts Institute of Technology (MIT) in collaboration with the Beth Israel Hospital (BIH), this database is a comprehensive collection of annotated electrocardiogram (ECG) signals, designed to support research and development in the area of cardiac arrhythmia analysis. Comprising recordings from 47 subjects, the database includes over 48 half-hour excerpts of two-channel ambulatory ECG recordings, obtained from a mixed population of inpatients (approximately 60%) and outpatients (approximately 40%) at the BIH Arrhythmia Laboratory between 1975 and 1979. The signals were digitized at a sampling rate of 360 samples per second per channel, with a resolution of 11 bits over a 10 mV range, ensuring high fidelity in the captured cardiac activities. One of the distinguishing features of the MIT-BIH Arrhythmia Database is its meticulous annotation. Each recording is accompanied by expert annotations for each heartbeat, identifying a wide range of arrhythmic and normal conditions. These annotations were made by well-trained observers and have been validated by multiple cardiologists, ensuring a high level of accuracy and reliability. The diversity of subjects and the variety of cardiac conditions represented in the dataset make it an invaluable resource for developing and testing algorithms intended for the detection and classification of arrhythmic events in ECG data.

To ensure a robust evaluation of our proposed model, we divided the total 90,460 samples data into training and test sets. In the interpatient scenario, we partitioned the dataset such that data from different patients were used for training and testing. Specifically, we used 80% of the data for training and 20% for testing. In the intrapatient scenario, we split the data from each patient into training and test sets. Again, we used an 80/20 split, where 80% of the data from each patient was used for training, and 20% was used for testing.

Before analysis, ECG signals from the MIT-BIH Arrhythmia Database undergo a normalization step to standardize their scale, crucial for the consistency and performance of our deep learning model. This process involves scaling the amplitude of the signals to a common range of 0 to 1, using the equation:

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where x_{norm} is the normalized signal, and $\min(x)$ and $\max(x)$ are the minimum and maximum values of the original signal x, respectively. This linear transformation ensures uniformity across the dataset, preparing the signals for effective feature extraction and analysis.

In evaluating the performance of our model for ECG signal classification, we employ several key metrics to comprehensively assess its accuracy, reliability, and applicability to real-world scenarios. These metrics are crucial for understanding how well the model can distinguish between different types of cardiac events, including Accuracy (Acc), Precision (Pre), Recall (Rec), and F1 score. Accuracy measures the overall correctness of the model across all classes. It is defined as the ratio of correctly predicted observations to the total observations. Recall (also known as Sensitivity) evaluates the model's ability to correctly identify positive instances for each

TABLE I
THE PERFORMANCE COMPARISON RESULTS WITH CURRENT ADVANCED MODELS BASED ON INTRAPATIENT PARTITIONING.

References	Arrhythmia Types	Acc (%)	Pre (%)	Rec (%)	F1 score (%)
Martis et al. [45]	5	93.48	98.31	99.27	98.79
Serkan et al. [46]	5	93.47	96.01	91.64	93.77
Acharya et al. [47]	5	95.14	66.56	96.94	78.93
Oh et al. [48]	5	97.88	97.26	98.50	97.88
Novotna et al. [49]	5	96.81	95.47	98.46	96.94
Wang et al. [26]	5	96.29	96.29	99.24	97.74
Fawaz et al. [50]	5	96.86	97.39	98.02	97.70
Talpur et al. [51]	5	96.49	96.84	99.16	97.99
Ours	5	**98.46**	**98.83**	**99.37**	**99.10**

TABLE II
THE PERFORMANCE COMPARISON RESULTS WITH CURRENT ADVANCED MODELS BASED ON INTERPATIENT PARTITIONING.

References	Arrhythmia Types	Acc (%)	Pre (%)	Rec (%)	F1 score (%)
Chazal et al. [52]	5	85.88	**66.0**	45.57	53.91
Tan and Le [53]	5	80.46	58.43	40.52	47.85
Wang et al. [26]	5	81.08	54.55	50.5	52.45
Novotna et al. [49]	5	82.81	65.32	46.43	54.28
Talpur et al. [51]	5	83.42	62.74	52.86	57.38
Chen et al. [54]	5	83.05	60.49	66.48	63.34
Ours	5	**86.84**	63.07	**70.11**	**66.4**

class, crucial for conditions where missing a positive case could be detrimental. Additionally, the Precision metric is employed to assess the proportion of positive identifications that were actually correct, essential in contexts where the cost of a false positive is high. Finally, to capture a balance between the sensitivity and precision of our model, especially in situations where there is a class imbalance, we calculate the F1 Score. This metric is the harmonic mean of precision and sensitivity, providing a single measure to assess the model's accuracy in identifying each class. These four metrics are defined below.

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

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

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

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where TP, TN, FP, and FN represent the true positives, true negatives, false positives, and false negatives, respectively. Employing these metrics allows for a nuanced evaluation of our model's performance in classifying ECG signals. It not only highlights the model's strengths in detecting various cardiac conditions but also identifies areas for improvement, guiding future enhancements to increase diagnostic accuracy and reliability.

B. Comparison with Other Methods

In the section dedicated to the performance comparison of our experiments, we embark on a detailed analysis to illustrate the effectiveness of our proposed model in classifying ECG signals. This comparative study is pivotal, as it not only benchmarks the capabilities of our model against existing advanced methods but also highlights the advancements our approach brings to the field of cardiac arrhythmia detection. To ensure a fair and comprehensive comparison, we select a range of models from the literature that have demonstrated proficiency in ECG signal classification. These models span various techniques, from traditional machine learning approaches to advanced deep learning architectures, providing a broad spectrum for evaluation, including [52], [45], [46], [47], [49], [50], [53], [54], [48], [26], [51], and our proposed method.

First, we introduce the classification of the dataset and the partitioning of experiments. According to the AAMI standard [55], five categories of arrhythmias are identified: Normal beats (N), Supra-ventricular ectopic beats (S), Ventricular ectopic beats (V), Fusion beats (F), and Unclassifiable beats (U). When this standard is applied to the MIT-BIH Arrhythmia Database, which classifies 15 types of heartbeats. Specifically, the N category includes Normal beat (Norm), Left bundle branch block beat (LBB), Right bundle branch block beat (RBB), Atrial escape beat (AE), and Nodal (junctional) escape beat (NE). The S category includes Atrial premature beat (AP), Aberrated atrial premature beat (aAP), Nodal (junctional) pre-

TABLE III
ABLATION EXPERIMENTS FOR MULTIMODAL DATA BASED ON THE INTERPATIENT PARTITIONING.

Data	Acc (%)	Pre (%)	Rec (%)	F1 score (%)
Signals	80.27	60.27	63.17	61.69
Images	76.32	56.34	57.86	57.09
Multimodal	**86.84**	**63.07**	**70.11**	**66.4**

TABLE IV
ABLATION EXPERIMENTS FOR MULTIMODAL DATA BASED ON THE INTRAPATIENT PARTITIONING.

Data	Acc (%)	Pre (%)	Rec (%)	F1 score (%)
Signals	91.75	92.64	94.56	93.59
Images	90.82	90.84	92.49	91.65
Multimodal	**98.46**	**98.83**	**99.37**	**99.10**

mature beat (NP), and Supraventricular premature or ectopic beat (SP). The V category includes Premature ventricular contraction (PVC) and Ventricular escape beat (VE). The F category includes the Fusion of ventricular and normal beat (fVN). The U category includes Paced beat (P), Fusion of paced and normal beat (fPN), and Unclassifiable beat (U). This classification encompasses a comprehensive range of cardiac rhythms, from standard to highly irregular patterns, showcasing the AAMI standard's application in categorizing diverse heartbeats found in the MIT-BIH Arrhythmia Database.

In the performance comparison of our study, we consider two distinct data partitioning scenarios: interpatient and intrapatient. These scenarios are crucial for evaluating the generalizability and adaptability of our model under different conditions, providing insights into its potential clinical applicability. Interpatient scenario involves partitioning the dataset such that the training and testing sets contain data from entirely separate groups of patients. No single patient's data is shared between these two sets. This scenario simulates a real-world application where the model is required to make predictions on data from new patients it has never encountered during training. The interpatient partitioning tests the model's ability to generalize across the physiological and pathological variations present in the broader population. In contrast, the intrapatient scenario involves partitioning the data from each patient into training and testing sets. This means that the model is trained and tested on different segments of data from the same patients. This scenario is particularly useful for assessing the model's performance in situations where longitudinal data is available for individuals, allowing us to understand how well the model can predict future cardiac events for a patient based on their past data. Both these data partitioning strategies offer valuable perspectives on the model's performance. Interpatient testing assesses the robustness of the model across diverse patient populations, a critical factor for clinical deployment. In contrast, intrapatient testing provides insight into the model's reliability and consistency in monitoring and diagnosing the same patient over time. Together, these scenarios ensure a comprehensive evaluation of our proposed model, highlighting its strengths and identifying areas for further refinement. Table I shows the comparison results with existing methods based on intrapatient partitioning. Table II shows the comparison results with existing methods based on interpatient partitioning. As can be seen from the results in the table, our proposed method surpasses all compared models in performance, which further demonstrates the effectiveness of the multi-modal feature fusion method proposed in this paper in ECG medical data analysis. Through this comparative analysis, we contribute to the ongoing dialogue in the biomedical signal processing community, advancing the collective understanding and development of effective diagnostic tools for cardiac arrhythmias. To show the details of each classification result, we show the results of confusion matrix results, as shown in Fig. 2.

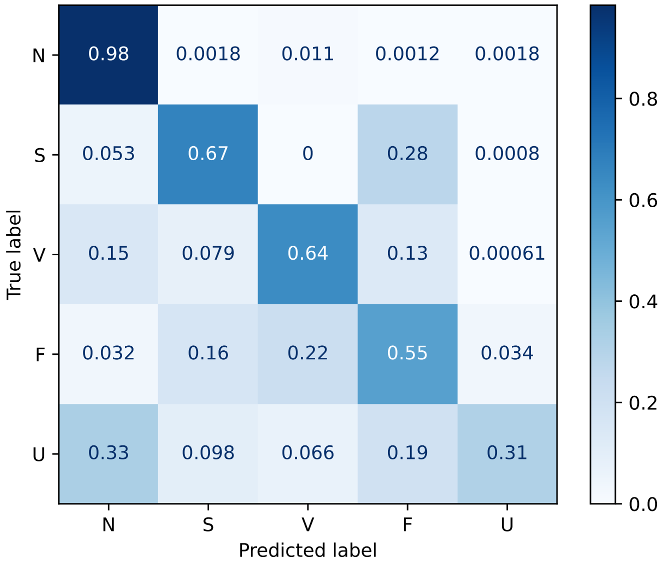

Fig. 2. The confusion matrix results based on the intrapatient data division.

C. Ablation study

In the experimental section dedicated to ablation studies, we aim to dissect the contributions of the multimodal feature fusion and the Neuro-Fuzzy modules within our proposed model, focusing on their roles in enhancing the task of ECG signal classification. These studies are designed to provide a deeper understanding of how each component influences the model's predictive performance, thereby validating the effectiveness of our integrated approach.

1) The Importance of Multimodal Features: In this section, our ablation experiments are structured to systematically evaluate the impact of the multimodal feature fusion. We compare the full model against variants where the multimodal feature fusion is removed separately. Specifically, we compare the experimental results in the following situations.

- Signals data: By training the model exclusively on ECG signal data, we assess the impact of omitting these spectral features. This variant tests the hypothesis that spectral image features, when fused with traditional ECG signal data, provide a significant boost in classification performance.

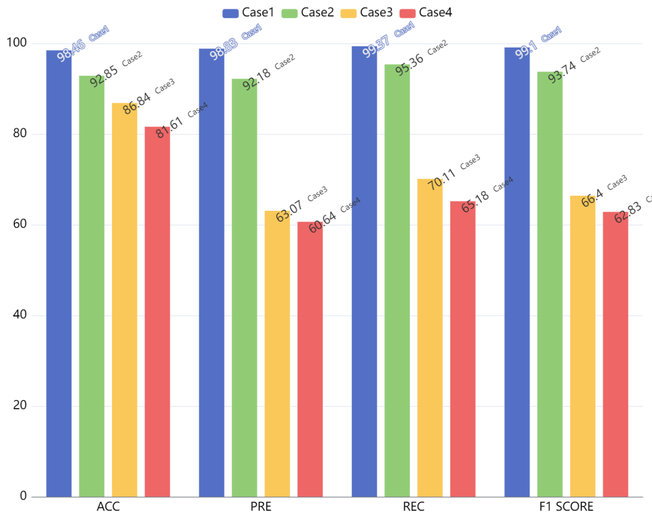

Fig. 3. The impact of the neuro-fuzzy module on model performance based on two data divisions.

- Images data: Likewise, by training the model exclusively on spectral images, we assess the impact of omitting ECG signal features. This variant tests the hypothesis that ECG signal data, when fused with traditional spectral image features, provide a significant boost in classification performance.
- Multimodal: Obviously, in this case, we use two kinds of data, namely ECG signal and spectral features, to jointly train and test the model.

Table III and Table IV depict the experimental results of data modality ablation in two cases of data partitioning respectively. Take Table III as an example, removing the spectral image features and relying solely on ECG signal data, the model's performance drops to an accuracy of 80.27%, the precision of 60.27%, the recall of 63.17%, and an F1 score of 61.69%. Likewise, using only image data, the model's performance drops to an accuracy of 76.32%, the precision of 56.34%, the recall of 57.86%, and an F1 score of 57.09%. The decline in performance metrics upon removing the multimodal feature fusion highlights the value of integrating spectral image features with ECG signal data. This fusion evidently enhances the model's ability to capture a more comprehensive representation of cardiac activities, significantly contributing to its predictive accuracy.

2) The Impact of Neuro-Fuzzy Model: In the proposed model, we use the neuro-fuzzy model to analyze the fused multi-modal features to achieve high-accuracy classification and prediction of ECG data. Here, we evaluate the contribution of the neuro-fuzzy module by replacing it with a standard deep learning classification layer (i.e., the fully connected layer). This allows us to determine the extent to which the fuzzy logic-based reasoning enhances the model's ability to handle uncertainties and improve classification accuracy. We compare the model performance in the following two situations:

- Case1: we train our proposed model in this paper with the neuro-fuzzy module based on the intrapatient data division.
- Case2: we train our proposed model in this paper without the neuro-fuzzy module based on the intrapatient data division.
- Case3: we train our proposed model in this paper with the neuro-fuzzy module based on the interpatient data division.
- Case4: we train our proposed model in this paper without the neuro-fuzzy module based on the interpatient data division.

Fig. 3 depicts the model performance in the above four cases on the four metrics. From the figure, we can clearly see that no matter which data division method is used, when the neuro-fuzzy module is removed, the performance of the model decreases. Similarly, the drop in performance metrics when the neuro-fuzzy module is replaced suggests that the fuzzy logic component plays a crucial role in handling the uncertainties and imprecisions inherent in ECG data. The neuro-fuzzy module improves the model's decision-making capabilities.

3) The Impact of Different Backbones: In our ablation study, we extend our examination to include the validation of various feature extraction models within our framework. Specifically, we investigate how replacing the Transformer-based feature extraction module with Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Residual Networks (ResNet), and DenseNet impacts the overall performance of our model in classifying ECG signals. The selected CNN, RNN, and ResNet models are representative of the state-of-the-art in deep learning approaches for ECG analysis, each known for their strengths in handling different aspects of signal processing. CNNs are renowned for their ability to capture local patterns through convolutional operations, RNNs excel in modeling sequential dependencies, and ResNets are effective in training deep networks by mitigating the vanishing gradient problem through residual connections. DenseNet introduces dense connections between layers, enhancing feature reuse and mitigating the vanishing gradient problem. Including DenseNet will allow us to evaluate its ability to capture detailed features in ECG signals. This comparative analysis aims to highlight the effectiveness of the Transformer in capturing the intricate patterns within ECG data as compared to other widely used architectures.

For this segment of our study, we crafted several model variants, each utilizing a different feature extraction module:

- CNN: A model where the Transformer is replaced with a CNN architecture, known for its ability to capture spatial hierarchies in data.
- RNN: This variant employs an RNN, particularly adept at handling sequential data, to process the ECG signals.
- ResNet: Incorporates a ResNet model for feature extraction, leveraging its deep residual learning framework to address vanishing gradients in deep architectures.
- DenseNet: Incorporates a EfficientNet model for feature extraction.

Each of these variants is evaluated under identical experimental conditions, ensuring a fair comparison. The models were assessed based on accuracy, sensitivity, specificity, precision,

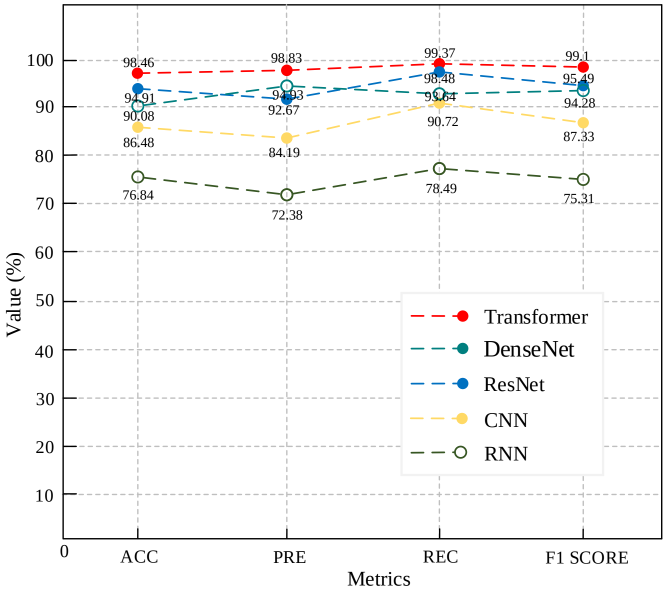

Fig. 4. The impact of different backbones on model performance based on the intrapatient data division.

and F1 score, mirroring the evaluation metrics applied to the full model. Fig. 4 depicts the impact of different backbones on model performance based on the intrapatient data division. As can be seen in the figure, the Transformer-based model outperformed the other variants, achieving the highest accuracy. While the ResNet variant showed competitive performance, the CNN and RNN variants lagged slightly behind, indicating their lesser capability in capturing the comprehensive features required for high accuracy in ECG classification. The Transformer model also led in precision and recall, crucial for minimizing false negatives and false positives, respectively. The RNN variant, although effective in handling sequential data, did not achieve the same level of performance, possibly due to its limitations in capturing long-term dependencies without complex modifications.

V. Conclusion

In summary, we introduce an innovative method for ECG big data analysis by combining deep learning techniques with neuro-fuzzy systems in this paper. Firstly, we transform ECG signals into spectrograms to obtain more information. Secondly, we extract features from one-dimensional ECG data signals and spectrum images based on the Transformer model. Thirdly, we employ a neuro-fuzzy model to analyze the fused multimodal features to achieve high-precision classification and prediction of ECG data. This research marks an advancement in the analysis of ECG big data and presents a solution that synergizes deep neuro-fuzzy models with multimodal feature fusion. The effectiveness of this integration is thoroughly evidenced through our experimental results, showcasing superior performance in ECG signal classification tasks. The ablation studies further illuminate the critical role of each component, particularly underscoring the value of multimodal feature fusion and the strategic advantage of employing a Transformer over traditional architectures like CNN, RNN, and ResNet. This approach not only improves the analysis's accuracy but also offers new insights and tools for processing and analyzing complex biomedical signals. However, there are also potential disadvantages to consider. The complexity of the Transformer-based model may lead to increased computational requirements, which could pose challenges for real-time applications or deployment in resource-constrained environments. Additionally, while the neuro-fuzzy system improves interpretability, it may also introduce additional complexity in the model training process.

References

[1] H. Blackburn, A. Keys, E. Simonson, P. Rautaharju, and S. Punsar, "The electrocardiogram in population studies: a classification system," *Circulation*, vol. 21, no. 6, pp. 1160–1175, 1960.

[2] N. Talpur, S. J. Abdulkadir, H. Alhussian, ·. M. H. Hasan, N. Aziz, and A. Bamhdi, "A comprehensive review of deep neuro-fuzzy system architectures and their optimization methods," *Neural Computing and Applications*, pp. 1–39, 2022.

[3] M. Yeganejou, S. Dick, and J. Miller, "Interpretable deep convolutional fuzzy classifier," *IEEE transactions on fuzzy systems*, vol. 28, no. 7, pp. 1407–1419, 2019.

[4] G. Srivastava, J. C.-W. Lin, D. Pamucar, and S. Kotsiantis, "Applications of fuzzy systems in data science and big data," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 1, pp. 1–3, 2020.

[5] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[6] S. A. Shufni and M. Y. Mashor, "Ecg signals classification based on discrete wavelet transform, time domain and frequency domain features," in *2015 2nd international conference on biomedical engineering (ICoBE)*. IEEE, 2015, pp. 1–6.

[7] Y. A. Altay and A. S. Kremlev, "Comparative analysis of ecg signal processing methods in the time-frequency domain," in *2018 IEEE conference of Russian young researchers in electrical and electronic engineering (EIConRus)*. IEEE, 2018, pp. 1058–1062.

[8] V. Mazaheri and H. Khodadadi, "Heart arrhythmia diagnosis based on the combination of morphological, frequency and nonlinear features of ecg signals and metaheuristic feature selection algorithm," *Expert Systems with Applications*, vol. 161, p. 113697, 2020.

[9] J. Wang, R. Li, R. Li, K. Li, H. Zeng, G. Xie, and L. Liu, "Adversarial de-noising of electrocardiogram," *Neurocomputing*, vol. 349, pp. 212–224, 2019.

[10] A. Biran and A. Jeremic, "Ecg based human identification using short time fourier transform and histograms of fiducial qrs features." in *BIOSIGNALS*, 2020, pp. 324–329.

[11] Z. Wu, T. Lan, C. Yang, and Z. Nie, "A novel method to detect multiple arrhythmias based on time-frequency analysis and convolutional neural networks," *IEEE Access*, vol. 7, pp. 170 820–170 830, 2019.

[12] S. S. Abdeldayem and T. Bourlai, "Automatically detecting arrhythmia-related irregular patterns using the temporal and spectro-temporal textures of ecg signals," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2301–2307.

[13] M. K. Gautam and V. K. Giri, "A neural network approach and wavelet analysis for ecg classification," in *2016 IEEE international conference on engineering and technology (ICETECH)*. IEEE, 2016, pp. 1136–1141.

[14] S. Pal and M. Mitra, "Detection of ecg characteristic points using multiresolution wavelet analysis based selective coefficient method," *Measurement*, vol. 43, no. 2, pp. 255–261, 2010.

[15] U. Desai, R. J. Martis, C. G. Nayak, K. Sarika, and G. Seshikala, "Machine intelligent diagnosis of ecg for arrhythmia classification using dwt, ica and svm techniques," in *2015 Annual IEEE India Conference (INDICON)*. IEEE, 2015, pp. 1–4.

[16] E. Alickovic and A. Subasi, "Medical decision support system for diagnosis of heart arrhythmia using dwt and random forests classifier," *Journal of medical systems*, vol. 40, no. 4, p. 108, 2016.

[17] S. Raj and K. C. Ray, "Sparse representation of ecg signals for automated recognition of cardiac arrhythmias," *Expert systems with applications*, vol. 105, pp. 49–64, 2018.

[18] Z. I. Attia, P. A. Noseworthy, F. Lopez-Jimenez, S. J. Asirvatham, A. J. Deshmukh, B. J. Gersh, R. E. Carter, X. Yao, A. A. Rabinstein, B. J. Erickson *et al.*, "An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction," *The Lancet*, vol. 394, no. 10201, pp. 861–867, 2019.

[19] H. Zhu, C. Cheng, H. Yin, X. Li, P. Zuo, J. Ding, F. Lin, J. Wang, B. Zhou, Y. Li *et al.*, "Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study," *The Lancet Digital Health*, vol. 2, no. 7, pp. e348–e357, 2020.

[20] Q. Xie, S. Tu, G. Wang, Y. Lian, and L. Xu, "Discrete biorthogonal wavelet transform based convolutional neural network for atrial fibrillation diagnosis from electrocardiogram." in *IJCAI*, 2020, pp. 4403–4409.

[21] P. Xiong, Y. Xue, M. Liu, H. Du, H. Wang, and X. Liu, "Detection of inferior myocardial infarction based on densely connected convolutional neural network," *Sheng wu yi xue Gong Cheng xue za zhi= Journal of Biomedical Engineering= Shengwu Yixue Gongchengxue Zazhi*, vol. 37, no. 1, pp. 142–149, 2020.

[22] S. Al-Zaiti, L. Besomi, Z. Bouzid, Z. Faramand, S. Frisch, C. Martin-Gill, R. Gregg, S. Saba, C. Callaway, and E. Sejdić, "Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram," *Nature communications*, vol. 11, no. 1, p. 3966, 2020.

[23] A. Mostayed, J. Luo, X. Shu, and W. Wee, "Classification of 12-lead ecg signals with bi-directional lstm network," *arXiv preprint arXiv:1811.02090*, 2018.

[24] 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," *Information Fusion*, vol. 53, pp. 174–182, 2020.

[25] S. Saadatnejad, M. Oveisi, and M. Hashemi, "Lstm-based ecg classification for continuous monitoring on personal wearable devices," *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 515–523, 2019.

[26] J. Wang, R. Li, R. Li, B. Fu, C. Xiao, and D. Z. Chen, "Towards interpretable arrhythmia classification with human-machine collaborative knowledge representation," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 7, pp. 2098–2109, 2020.

[27] X. Xu, S. Jeong, and J. Li, "Interpretation of electrocardiogram (ecg) rhythm by combined cnn and bilstm," *Ieee Access*, vol. 8, pp. 125 380–125 388, 2020.

[28] Y. Wang, G. Yang, S. Li, Y. Li, L. He, and D. Liu, "Arrhythmia classification algorithm based on multi-head self-attention mechanism," *Biomedical Signal Processing and Control*, vol. 79, p. 104206, 2023.

[29] Y. Zhang, J. Yi, A. Chen, and L. Cheng, "Cardiac arrhythmia classification by time–frequency features inputted to the designed convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 79, p. 104224, 2023.

[30] L. Subramanyan and U. Ganesan, "A novel deep neural network for detection of atrial fibrillation using ecg signals," *Knowledge-Based Systems*, vol. 258, p. 109926, 2022.

[31] S. Śmigiel, K. Pałczyński, and D. Ledziński, "Ecg signal classification using deep learning techniques based on the ptb-xl dataset," *Entropy*, vol. 23, no. 9, p. 1121, 2021.

[32] A. Rath, D. Mishra, G. Panda, S. C. Satapathy, and K. Xia, "Improved heart disease detection from ecg signal using deep learning based ensemble model," *Sustainable Computing: Informatics and Systems*, vol. 35, p. 100732, 2022.

[33] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional lstm network model for ecg signal classification," *Computers in biology and medicine*, vol. 96, pp. 189–202, 2018.

[34] 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," *Information Fusion*, vol. 53, pp. 174–182, 2020.

[35] G. Bortolan and W. Pedrycz, "Fuzzy descriptive models: an interactive framework of information granulation [ecg data]," *IEEE Transactions on fuzzy Systems*, vol. 10, no. 6, pp. 743–755, 2002.

[36] R. Xiao, C. Ding, X. Hu, G. D. Clifford, D. W. Wright, A. J. Shah, S. Al-Zaiti, and J. K. Zègre-Hemsey, "Integrating multimodal information in machine learning for classifying acute myocardial infarction," *Physiological Measurement*, vol. 44, no. 4, p. 044002, 2023.

[37] R. Sharma, C. F. Eick, and N. V. Tsekos, "Sm2n2: A stacked architecture for multimodal data and its application to myocardial infarction detection," in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11*. Springer, 2021, pp. 342–350.

[38] M. R. Schwob, A. Dempsey, F. Zhan, J. Zhan, and A. Mehmood, "Robust multimodal heartbeat detection using hybrid neural networks," *IEEE Access*, vol. 8, pp. 82 201–82 214, 2020.

[39] M. Zarrabi, H. Parsaei, R. Boostani, A. Zare, Z. Dorfeshan, K. Zarrabi, and J. Kojuri, "A system for accurately predicting the risk of myocardial infarction using pcg, ecg and clinical features," *Biomedical Engineering: Applications, Basis and Communications*, vol. 29, no. 03, p. 1750023, 2017.

[40] R. Xiao, C. Ding, X. Hu, and J. Zègre-Hemsey, "Ml for mi-integrating multimodal information in machine learning for predicting acute myocardial infarction," *medRxiv*, pp. 2022–10, 2022.

[41] Z. Ahmad, A. Tabassum, L. Guan, and N. M. Khan, "Ecg heartbeat classification using multimodal fusion," *IEEE Access*, vol. 9, pp. 100 615–100 626, 2021.

[42] T. Phan, D. Le, P. Brijesh, D. Adjeroh, J. Wu, M. O. Jensen, and N. Le, "Multimodality multi-lead ecg arrhythmia classification using self-supervised learning," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2022, pp. 01–04.

[43] 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.

[44] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[45] R. J. Martis, U. R. Acharya, K. Mandana, A. K. Ray, and C. Chakraborty, "Cardiac decision making using higher order spectra," *Biomedical Signal Processing and Control*, vol. 8, no. 2, pp. 193–203, 2013.

[46] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. San Tan, "A deep convolutional neural network model to classify heartbeats," *Computers in Biology and Medicine*, vol. 89, pp. 389–396, 2017.

[47] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2015.

[48] S. L. Oh, E. Y. Ng, R. San Tan, and U. R. Acharya, "Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats," *Computers in Biology and Medicine*, vol. 102, pp. 278–287, 2018.

[49] P. Novotna, T. Vicar, M. Ronzhina, J. Hejc, and J. Kolarova, "Deep-learning premature contraction localization in 12-lead ecg from whole signal annotations," in *2020 Computing in Cardiology*. IEEE, 2020, pp. 1–4.

[50] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[51] N. Talpur, S. J. Abdulkadir, and M. H. Hasan, "A deep learning based neuro-fuzzy approach for solving classification problems," in *2020 International Conference on Computational Intelligence (ICCI)*. IEEE, 2020, pp. 167–172.

[52] P. De Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1196–1206, 2004.

[53] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[54] L. Chen, C. Lian, Z. Zeng, B. Xu, and Y. Su, "Cross-modal multiscale multi-instance learning for long-term ecg classification," *Information Sciences*, vol. 643, p. 119230, 2023.

[55] ANSI-AAMI, "Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms," *Association for the Advancement of Medical Instrumentation*, 1998.