Artificial Intelligence Review (2025) 58:261
https://doi.org/10.1007/s10462-025-11259-x

A survey of transformers and large language models for ECG diagnosis: advances, challenges, and future directions

Mohammed Yusuf Ansari[1,2] · Mohammed Yaqoob[1] · Mohammed Ishaq[1] · Eduardo Feo Flushing[2] · Iffa Afsa changaai Mangalote[3] · Sarada Prasad Dakua[3] · Omar Aboumarzouk[3] · Raffaella Righetti[1] · Marwa Qaraqe[4]

Accepted: 11 May 2025 / Published online: 4 June 2025
© The Author(s) 2025

Abstract
Electrocardiograms (ECGs) are widely utilized in clinical practice as a non-invasive diagnostic tool for detecting cardiovascular diseases. Convolutional neural networks (CNNs) have been the primary choice for ECG analysis due to their capability to process raw signals. However, their localized convolutional operations limit the ability to capture long-range temporal dependencies across heartbeats, impeding a comprehensive cardiovascular assessment. To address these limitations, transformer-based frameworks have been introduced, employing self-attention mechanisms to effectively model complex temporal patterns over entire ECG sequences. Recent advancements in large language models (LLMs) have further expanded the utility of transformers by enabling multimodal integration and facilitating zero-shot diagnosis, thereby enhancing the scope of ECG-based clinical applications. Despite the increasing adoption of these methodologies, a comprehensive survey systematically examining transformer and LLM-based approaches for ECG analysis is absent from the literature. Consequently, this article surveys existing methods and proposes a novel hierarchical taxonomy based on the complexity of diagnosis, ranging from single-beat analysis to multi-beat and full-length signal evaluations. A thorough cross-category comparison is performed to highlight overarching commonalities and limitations. In light of these limitations, the paper presents a discussion of critical gaps and introduces new future directions aimed at improving ECG representation, enhancing positional encodings, refining self-attention architectures, and addressing challenges related to hallucinations and confidence measures in LLMs. The insights and guidelines presented aim to inform future research and clinical practices, enabling the next generation of intelligent ECG diagnostic systems.

Keywords Arrhythmia · ECG representation · Hallucination · Large language models · Myocardial infarction · Positional encoding · Self-attention architecture · Single-beat and multi-beat analysis · Sleep apnea · Zero-shot diagnosis

M.Y. Ansari, M. Yaqoob, and M. Ishaq have been contributed equally to this work.

Extended author information available on the last page of the article

1 Introduction

An electrocardiogram (ECG) is a widely adopted multi-channel signal that records the cardiac electrical activity during the contraction and relaxation phases of the atria and ventricles. ECGs are a significant first step in several cardiac-related clinical protocols. To elaborate, ECGs are promptly performed in emergencies to assess cardiac function and guide immediate treatment in cases such as chest pain or suspected myocardial infarction (Gustafsson et al. 2022). ECGs are also integral to routine physical exams and preoperative assessments, especially for patients with cardiovascular risk factors, to detect asymptomatic abnormalities and evaluate cardiac risk before anesthesia and surgery (van Klei et al. 2007). Additionally, implantable loop recorders (Milstein et al. 2020) and wearable ECG devices (Bouzid et al. 2022) enable longitudinal ECG monitoring (Ha et al. 2021) to track disease progression and assess the effectiveness of cardiac treatments. Consequently, ECG serves as a primary cardiac monitoring test in healthcare settings because of its non-invasive nature, cost-effectiveness, broad availability, versatility, reliability, and ability to provide immediate and valuable insights into cardiac health.

Conventionally, electrophysiologists perform ECG analysis by extracting ECG features (i.e., amplitudes and duration of ECG waves). Subsequently, the derived features are compared with the clinically established ECG normal values for the subject's age group to formulate a diagnosis (Dickinson 2005; Rijnbeek et al. 2014). However, the manual nature of conventional ECG analysis raises several challenges. Specifically, manual ECG analysis is subjective, which can be time-consuming and cause human error, potentially leading to variability in diagnosis and missed subtle abnormalities. Machine learning addresses these limitations by processing handcrafted features to provide objective and instantaneous diagnosis of cardiac conditions. To elucidate, decision trees recursively partition the ECG features into subsets generating tree-like interpretable structures, facilitating ECG diagnosis (Kumari and Sai 2022). Bagging for ECG diagnosis employs multiple weak learners (e.g., decision trees) trained on a subset of data and aggregates (e.g., voting) their output for enhanced diagnostic accuracy (Mert et al. 2014). Similarly, boosting employs a strategy to train multiple weak learners sequentially, where each model tries to minimize the errors of its predecessor by placing higher emphasis on misdiagnosed ECG instances (Shi et al. 2019). Altogether, machine learning strategies have significantly advanced automated ECG diagnosis but lack the ability to leverage raw ECG data, rely heavily on ECG domain knowledge for feature extraction, and fail to utilize subtle ECG features that are not apparent or known to electrophysiologists. Deep learning has addressed these challenges by learning mappings between raw ECG signals and diagnostic tasks (i.e., representation learning) without the need for manual feature extraction. Predominantly, CNNs have become the standard for ECG analysis and diagnosis. The convolutional layers autonomously extract task-relevant features, which are then passed through a multilayer perceptron to predict ECG disease diagnosis (Rashed-Al-Mahfuz et al. 2021; Makimoto et al. 2020; Karthiga et al. 2022).

ECG data is inherently sequential and captures heartbeats over time steps. Thus, identifying and tracking subtle changes across multiple beats in the wave morphology, amplitudes, and durations could play a significant role in diagnosing rhythmic abnormalities (e.g., arrhythmia) among other cardiac diseases (e.g., long QT syndrome and myocardial ischemia). The mathematical formulation of CNNs using fixed kernel sizes intrinsically enables them to learn local spatial correlations. As a result, CNNs lack the ability to capture long-range dependencies due to their localized receptive fields and reliance on pooling layers. Transformer architectures overcome this limitation using self-attention mechanisms that capture long-range dependencies by attending to all positions in the sequence simultaneously. Conventionally, transformers are used for natural language tasks, allowing them to learn the context/relevance of the word in a sentence with respect to other words in the current and previous sentences, improving the performance in language translation and generation tasks. Given the success of transformers for processing temporal data (e.g., language and speech), transformer-based networks with vanilla/modified self-attention, and hybrid CNN-transformer networks have been gaining traction in automated ECG diagnosis literature. Particularly, atypical heartbeat (Peng et al. 2024), abnormal cardiac rhythm (Ji et al. 2024), obstructive sleep apnea (Wang et al. 2024), and myocardial infarction (Liu et al. 2024) diagnosis have significantly benefited from transformer-based neural network methodologies.

LLMs have recently advanced ECG diagnosis and analysis by leveraging their increased parameter count and deep layers, which enable them to capture complex patterns and nuanced relationships in ECG signals. LLMs (e.g., GPT) are typically pre-trained using large-scale datasets on a generic natural language task and subsequently fine-tuned for ECG diagnosis and analysis tasks. A unique attribute of LLMs is their ability to handle multi-modal data, such as ECG signals, patient history in natural language, and structured demographic information, to enhance diagnostic Sensitivity (Qiu et al. 2023). Additionally, the comprehensive language understanding capabilities of LLMs can be harnessed by integrating them with diagnostic databases, enabling zero-shot diagnoses. Recent advances in LLMs have enabled the integration of diverse clinical data, including ECG signals, radiology images, and corresponding reports, to support comprehensive patient evaluations. For example, Thapa et al. (2024) introduced MoRE, which fuses X-ray images, ECG signals, and cardiology reports into a unified representation for comprehensive patient evaluation. Trained with Lora-Peft and optimized using contrastive loss to align modality-specific features, MoRE enables effective multimodal retrieval and zero-shot classification, achieving state-of-the-art performance on downstream task datasets such as Mimic-IV, CheXpert, Edema Severity, and PTB-XL. Similarly, Guo et al. (2024) developed ECGChat, a multimodal LLM that addresses discrepancies between ECG waveforms and textual cardiology reports. ECGChat supports zero-shot report retrieval and the autonomous generation of detailed ECG analyses by employing contrastive learning to synchronize ECG data with corresponding reports. Recently, Yu et al. (2023) proposed a zero-shot retrieval-augmented diagnosis technique, where LLMs retrieve ECG expert knowledge from a curated database to analyze ECG data without needing extensive prior training for diagnosing arrhythmia and sleep apnea. LLMs are currently being used in real-world settings to generate discharge summary reports (Chua et al. 2024), with pilot studies on RUSSEL GPT. Additionally, randomized clinical trials are being conducted to evaluate their impact on diagnostic reasoning, demonstrating their potential to influence clinical decision-making (Gallifant et al. 2025).

ECG analysis for diagnosis has been a constantly evolving domain, continuously adapting to the latest advancements in mainstream machine learning and deep learning. Several surveys have been published to capture the progress in automated ECG diagnosis research. These surveys focus on specific cardiac diseases, such as myocardial infarction (Han et al. 2024), hypertrophic cardiomayopathy (Maron et al. 2022), arrhythmia (Ebrahimi et al. 2020), and long QT syndrome (Dehkordi et al. 2024). Reviews have also focused on the use of specific computational approaches, like machine learning (Salari et al. 2022), CNNs (Liu et al. 2021), and unsupervised learning (Nezamabadi et al. 2022).

To our knowledge, no existing work comprehensively captures the progress and scope of transformer-based methodologies and the application of LLMs in the context of automated ECG diagnosis. As such, this survey comprehensively reviews transformer-based methodologies and their evolution into LLMs for ECG diagnosis, highlighting their technical novelty, application, limitations, and essential future directions. Particularly, the core contributions of this review are as follows:

- The categorization of the reviewed method follows a hierarchical analysis, starting with granular abnormal beat detection and advancing to more complex rhythm-level analysis within arrhythmias, before extending to other diseases, such as sleep apnea and cardiovascular diseases (CVDs) (e.g., myocardial infarction), among others. The review also captures the evolution of the transformer-based methodologies into LLMs, highlighting their applications in ECG-based diagnosis.
- Tabular summaries are provided to comprehensively capture ECG dataset details, methodological innovations, core results, and notable observations for each method discussed in the review. Additionally, the textual descriptions offer insights complementing the tabular summaries, providing descriptions and limitations specific to each work.
- The review conducts a cross-category comparison of methods to uncover overarching commonalities and limitations in transformer and LLM approaches for ECG-based diagnosis.
- In light of the limitations, the survey proposes essential ECG embedding enhancements and architectural innovations within the transformer and LLM pipelines. Subsequently, the review introduces strategies to overcome the critical shortcomings of LLMs in ECG diagnosis, potentially enhancing their transparency.

For transparency, this review adheres to the following inclusion and exclusion criteria: The manuscript only includes articles from peer-reviewed journals and conferences that have undergone a robust peer-review process for the last five years (i.e., 2019-2024). Particularly, the Google Scholar search engine was used in July-August 2024 to gather the articles within the scope of this review. Specific search queries utilized are as follows: "ECG Arrhythmia Transformers", "ECG Analysis Transformers", "ECG Diagnosis Transformers", "ECG Sleep Apnea Transformers", "ECG diagnosis LLMs", and "ECG Analysis Large Language Models". Popular shortlisted articles were cross-referenced to discover additional articles within the scope of this review.

The remainder of this survey is structured as follows: Sect. 2 provides the fundamental conceptual understanding of ECG and its role in CVD diagnosis along with an overview of transformer architecture for ECG analysis. Section 3 comprehensively reviews the transformer and LLM methodologies for ECG-based diagnosis, highlighting commonalities and limitations across these methods. Section 4 presents novel technical innovations to further improve the state-of-the-art for ECG-based diagnosis and highlights novel applications of LLMs. Finally, sect. 5 summarizes the findings and concludes the paper.

2 ECG and transformer fundamentals

The section provides an overview of essential concepts pertaining to ECGs and transformer architectures that are necessary to understand the reviewed articles.

2.1 ECG leads and waveform basics

A standard ECG for adults is a 12-lead multi-channel signal captured using 10 electrodes (Ansari et al. 2024). As shown in Fig. 1, the 12 leads of the ECG can be broadly classified into limb leads (i.e., I, II, III, aVL, aVR, and aVF) and chest leads (i.e., V1-V6). Six electrodes are placed on the precordial region and capture their corresponding chest leads. The remaining four electrodes are placed on the limbs, with one of them (typically right leg) serving as the ground/reference electrode. Limb leads measure the electrical activity in the frontal plane of the body, providing insights into the heart's rhythm and axis. Specifically, leads I, II, and III are standard limb leads that are captured by taking the difference of electrodes on the right arm, left arm, and left leg. Leads aVL, aVR, and aVF are augmented limb leads calculated by taking the average of two limb electrodes with respect to the third limb electrode. Key views covered by the limb leads include the lateral and inferior walls of the heart. In contrast, the chest leads measure the electrical activity in the horizontal plane of the body, offering detailed views of the heart's anterior and lateral walls. Specifically, leads V1 and V2 are oriented toward the right ventricle, offering a direct view of the septum wall. Leads V3 and V4 are positioned to face the interventricular septum anteriorly, capturing the electrical signals from the muscle that separates the left and right ventricles. Finally, leads V5 and V6 are directed toward the left ventricle and focus on the lateral wall (Wasimuddin et al. 2020; Ansari et al. 2023).

ECG waveform corresponds to the contraction and relaxation of the cardiac chambers, enabling blood flow. Specifically, the P-wave captures atrial depolarization that occurs with the generation of electrical impulse in the sinoatrial node (i.e., natural cardiac pacemaker), causing the atria to contract and push blood into the ventricles. Abnormalities in P-wave morphology and duration could indicate atrial enlargement (Yokota et al. 2021), atrial fibrillation (Rasmussen et al. 2020), or other atrial abnormalities. The QRS complex represents

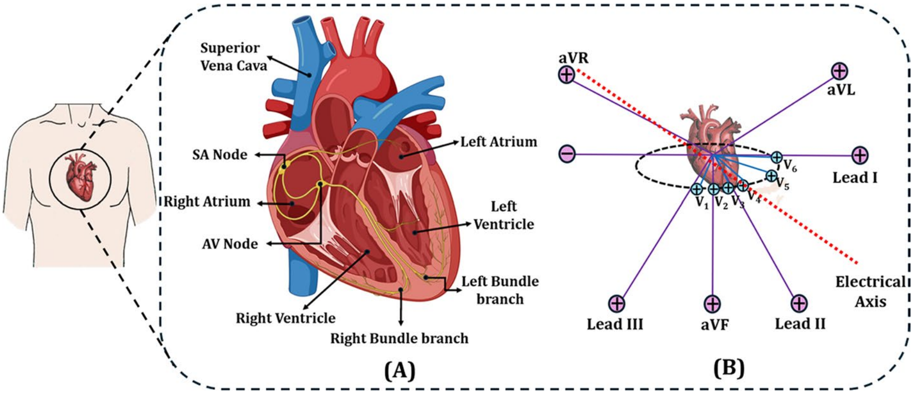

Fig. 1 Cardiac anatomy and Electrocardiogram lead placement. **A** Cross-sectional view of the heart illustrating key structures of the cardiac conduction system. **B** Standard 12-lead ECG electrode positions

ventricular depolarization, which occurs as the electrical impulse travels through the ventricles, pumping blood to the rest of the body. QRS waveform is crucial for assessing ventricular function and extended QRS duration or abnormal morphology (e.g., bundle branch blocks) can indicate conduction delays (Wu et al. 2021), ventricular hypertrophy (Pelliccia et al. 2023), or myocardial infarction (Luo et al. 2020). Finally, the T-wave encapsulates the process of the ventricles returning to their resting state after contraction. Abnormalities such as T wave inversion, flattening, or peaking can indicate ischemia (Li et al. 2021) or cardiomyopathies (D'Ascenzi et al. 2020). The duration of the individual waves and the intervals between them hold significant diagnostic importance. The PR interval is the duration from the onset of the P wave to the start of the QRS complex, capturing the time taken for the electrical impulse to travel from the atria to the ventricles. The ST segment represents the period between the end of ventricular depolarization and the beginning of ventricular repolarization. The QT interval indicates the net time for ventricular depolarization and repolarization, from the start of the QRS complex to the end of the T wave.

2.2 Overview of transformers and LLMs

Deep neural network-based approaches have gained significant momentum across a range of domains, including natural language processing (Gillioz et al. 2020; Tetko et al. 2020), geoscience (Yaqoob et al. 2024, 2025; Yaqoob et al. 2025a; Dahmani et al. 2025; Yaqoob et al. 2025b), speech recognition (Dong et al. 2018; Karita et al. 2019), medical imaging (Ansari et al. 2023; Shamshad et al. 2023; Ansari et al. 2024; Li et al. 2023; Ansari et al. 2022; Akhtar et al. 2021; Ansari et al. 2022; Ansari and Qaraqe 2023), and biomedical signal analysis (Zhang et al. 2023; Lih et al. 2023; Afsa et al. 2024). Recently, transformer architecture has been proposed to overcome the design limitations of previous sequence analysis models such as Recurrent Neural Networks (RNNs) (Salloum and Kuo 2017) and Long Short-Term Memory networks (LSTMs) (Hou et al. 2019). RNNs/LSTMs lack the ability to capture long-range dependencies due to issues like vanishing gradients and fixed-length context, limiting their understanding of how a data point relates to other points in a long sequence. Furthermore, the sequential nature of these models leads to longer training times, limiting their efficiency, especially with large datasets. Transformers overcome these challenges by utilizing self-attention mechanisms that allow for parallel processing of all elements in a sequence (see Fig. 2 for the schematic architecture of the transformer). This parallelism addresses the inefficiencies of sequential processing and enables the model to maintain a more comprehensive contextual understanding across entire sequences. Specific components that are combined to form the transformer architecture for ECG analysis are as follows:

2.2.1 Component 1: data embeddings

Transformers were initially developed for natural language processing (NLP) tasks, with one primary task being predicting the next word in a sequence. However, raw textual data (e.g., words) cannot be directly processed by the transformer, as it is designed to operate on numerical inputs. To overcome this challenge, words are represented as a dense vector in a high-dimensional space (i.e., embeddings) (Takase and Kobayashi 2020). Embeddings are carefully designed/learned to capture nuance details from the data. For instance, embeddings

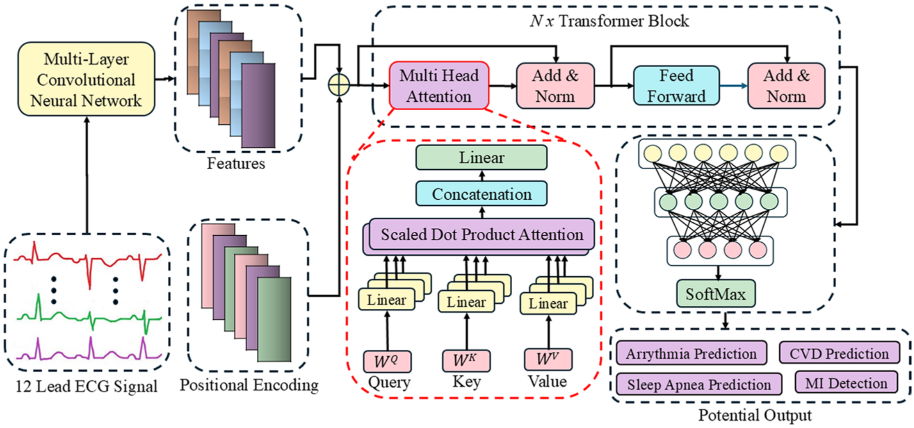

Fig. 2 Design of a transformer-based architecture for ECG analysis and cardiovascular disease prediction

capture semantic similarity with similar words with related meanings positioned closely in the embedding space. Additionally, advanced embeddings are context-aware, meaning that the vector representation of words is adjusted based on the surrounding context (e.g., the words "bank" may have different meanings in the context of river and finance), thus capturing nuanced meanings. In the context of ECG, embeddings are typically learned from ECG leads through a sequence of convolutional blocks (i.e., representation learning), enabling the network to learn condensed and meaningful representations of ECG signals for the given task.

2.2.2 Component 2: positional encoding

Transformers process all data points of a sequence in parallel, which enhances efficiency and the ability to capture long-range dependencies. However, this parallelism means that the embeddings lack intrinsic information about the order of the sequence. To address this, positional encodings are generated and added to the word embeddings, providing the transformer with the necessary information about the position of each data point within the sequence (Su et al. 2024). In NLP and ECG diagnosis tasks, it is common to employ positional embeddings that combine sine and cosine functions. For each position, a positional encoding vector is generated by applying the sine function to even-numbered dimensions and the cosine function to odd-numbered dimensions. Additionally, the frequency of the sine and cosine waves varies with the position in the sequence. Given a fixed position in the embedding vector, as the position in sequence increases, the frequency increases, thus, creating a unique pattern for each position. It is crucial to note that positional encodings have small values relative to the values in the word embeddings, ensuring that the addition of positional encodings doesn't overwhelm or dominate the information in the word embeddings. Mathematically, the sine-cosine positional encoding can be expressed as:

$$PE_{(pos,i)} = \begin{cases} \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) & \text{if } i \text{ is even} \\ \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) & \text{if } i \text{ is odd} \end{cases} \tag{1}$$

Here, *pos* represents the data point position in the sequence, *i* represents the index within the embedding vector, and *d* is the dimensionality of the embeddings. $10000^{\frac{2i}{d}}$ scales the input to create distinct and smooth positional encodings across the dimensions.

2.2.3 Component 3: self-attention

The fundamental novelty of the transformer architecture is the self-attention mechanism (Vaswani et al. 2017). Specifically, given a condensed representation of ECG leads, self-attention computes the relevance of a feature, such as the P-wave amplitude, in relation to other critical features like the R-peak and T-wave amplitude within the same or across different beats. To achieve this, the model linearly transforms the ECG embeddings into Query (Q), Key (K), and Value (V) vectors. Mathematically, the vectors are generated as follows:

$$\mathrm{Q} = \mathrm{XW}_Q, \quad \mathrm{K} = \mathrm{XW}_K, \quad \mathrm{V} = \mathrm{XW}_V \tag{2}$$

Here, X represents the ECG embeddings. The matrices W_Q, W_K, and W_V are learnable weight matrices that project the input embeddings into three distinct spaces to generate the Q, K, and V matrices. The Query vector represents the feature whose influence or relevance the model seeks to understand (e.g., the P-wave amplitude). The Key vectors encode related ECG features (e.g., R-peak, T-wave) to determine their relevance to the Query. The Value vector contains the actual ECG data that is weighted according to the attention scores derived from the Q-K interactions. Mathematically, the attention scores are computed as:

$$\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^\top}{\sqrt{d_k}}\right)\mathrm{V} \tag{3}$$

Here, d_k represents the dimensionality of keys, and $\sqrt{d_k}$ scales the scores to prevent large scores during training. The resulting weighted sum is a contextually enriched representation of the initial ECG embeddings that enhances the model's ability to analyze complex inter-beat and intra-beat relationships.

2.2.4 Component 4: multi-head attention

Multiple self-attention mechanisms in parallel can allow the transformer model to attend to different aspects of the ECG signal (Voita et al. 2019). This idea is implemented in transformers using multiple heads (i.e., multi-head self-attention (MHSA)) parallelly. To elaborate, one head can identify the relationships between P-wave features across multiple beats, which is essential for analyzing atrial activity. Another head could concentrate on the timing and amplitude of R-peaks, evaluating heart rhythm and ventricular depolarization. MHSA can be implemented by first splitting the Q, K, and V matrices (equation 2) along the embedding dimension, enabling each head to operate on a subspace of these matrices with dimension $d_k = d_{model}/h$. Subsequently, each head can compute the attention score from $\mathbf{Q_h}$ and $\mathbf{K_h}$ and reweigh $\mathbf{V_h}$ as shown in equation 3. The heads are then combined to generate the output. Mathematically, this can be represented as:

$$\mathrm{Output} = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h)\mathrm{W}_O \tag{4}$$

Here, W_O represents the linear transformation used for generating the output of MHSA.

2.2.5 Component 5: add and norm and feed-forward networks

After the MHSA reweighs the input ECG embeddings, the feature maps go through the Add&Norm layer. The "Add" component represents a residual connection, allowing the model to cope with vanishing gradients and dilution of feature information that may occur during the MHSA transformations. Next, the "Norm" depicts layer normalization which normalizes the feature map by using the sum across the channel dimension. Layer normalization minimizes the internal covariate shift that may happen during training and speeds up the convergence of the model. Subsequently, a series of fully connected layers with activation functions are used to introduce non-linearity to the model. The fully connected layers are arranged in an inverted bottleneck layout. To elaborate, the first linear layer expands the dimensionality of the input (e.g., $4 \times d_{model}$), and the second layer reduces it back to the original size (i.e., d_{model}) (Geva et al. 2020). The expansion allows the transformer model to utilize higher dimensional feature space, enabling the capture of detailed and abstract features before compressing to the initial embedding space. The Add&Norm process is repeated after the feature map passes through the fully connected layers.

3 Reviewed methods

This section presents the findings from the surveyed literature, focusing on the role of transformers and LLMs for ECG-based disease classification. The review is organized hierarchically, beginning with granular abnormal beat detection and rhythm-level analysis within arrhythmias, then extending to other diseases, such as sleep apnea and CVDs (see Fig. 3 for a timeline of relevant studies). The section also covers the recent applications of LLMs for ECG analysis and diagnosis.

3.1 Arrhythmia

3.1.1 Beat classification

Transformer-based architectures have gained prominence in classifying abnormal heartbeats from ECG signals. Several studies focus exclusively on employing transformers to process raw ECG signals, leveraging their ability to capture both temporal and spatial features through enhancements like convolutional layers and sequential processing units (refer to Table 1). El-Ghaish and Eldele (2024) present ECGTransForm by integrating a bidirectional transformer (BiTrans) mechanism with multi-scale convolutions and a channel re-calibration module (CRM). The architecture differs from standard transformer models by incorporating multi-scale convolutional layers that capture spatial features across different scales, followed by a CRM that recalibrates channel-wise features to enhance inter-dependencies between channels. The Bidirectional transformer captures temporal dependencies from both past and future contexts. However, a limitation of the paper is the potential computational complexity introduced by the Bidirectional transformer. Hu et al. (2021) propose a robust wave-feature adaptive heartbeat classification using transformers. An adaptive

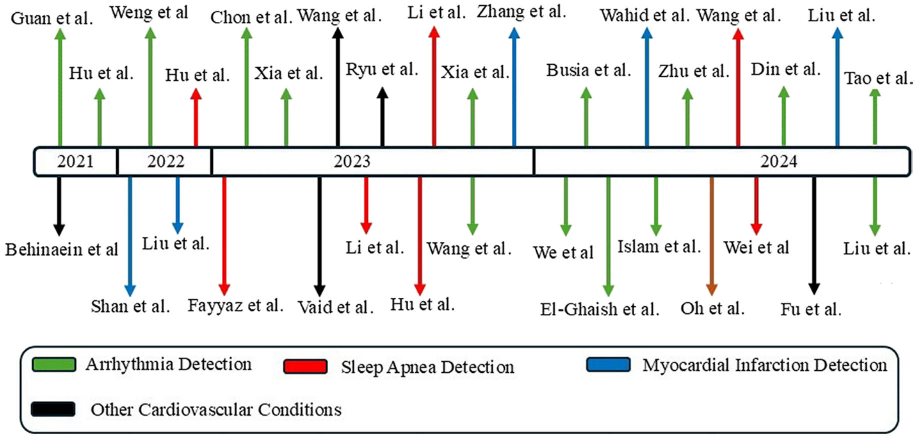

Fig. 3 Timeline of studies utilizing transformer-based architectures for ECG analysis in various cardiovascular conditions, categorized by arrhythmia detection, sleep apnea detection, myocardial infarction detection, and other cardiovascular conditions

heartbeat segmentation method is proposed, which dynamically adjusts the segmentation window based on the local RR interval. This modification allows the model to focus on relevant heartbeat segments within the ECG waveform. The architecture employs a standard transformer encoder that receives ECG wave features produced by 1D convolutional layers. However, a limitation of this approach is its reliance on the accuracy of the RR interval calculation for segmentation; any error in RR interval detection can propagate through the model. Islam et al. (2024) propose CAT-Net architecture. CAT-Net comprises CNN layers that capture the morphological features of the ECG signals, while the MHSA refines these features by reweighing the relevant segments. However, this model's evaluation on balanced datasets remains a limitation, as its performance on minority classes is sub-optimal. Busia et al. (2024) present a novel "Tiny Transformer" model optimized for low-power arrhythmia classification on microcontrollers, achieving a balance between high accuracy (98.97% on 8-bit quantization) and low computational complexity (requiring only 6,649 parameters and 0.09mJ per inference). The architecture utilizes a standard ViT block modified for 1D ECG signals with a single standard encoder block. A convolutional embedding stage prepares the ECG signal patches for processing with ViT. The authors limit the embedding size to 16 and the number of heads to 8 to minimize computational cost. Despite these innovations, the model's dependency on a single MHSA limits its ability to fully capture complex interactions within ECG signals, potentially impacting its generalizability for noisy ECG signals.

Liu et al. (2024) introduce the PSC-Net architecture. PSC-Net uses a modified transformer encoder where traditional fully connected layers are replaced with gated recurrent unit (GRU) layers that improve the model's ability to handle sequential data by reducing the risk of information loss across layers. The architecture also includes a unique feature fusion block that combines local features extracted by the GLNet (a CNN-based module) with global features obtained from the transformer encoder. However, a limitation of this approach is the potential overfitting caused by the extensive use of feature fusion and weighted residual connections. Lastly, Din et al. (2024) implements a hybrid model that integrates CNN, LSTM, and transformer in a single architecture. The proposed model capi-

Table 1 Summary of transformer-based models using raw ECG signals for beat classification

Reference (Year)	Dataset	Core methodology	Results	Remarks
El-Ghaish and Eldele (2024)	MIT-BIH (Moody and Mark 2001) + PTB (Bousseljot et al. 1995)	ECGTransForm	Accuracy: 99.35±0.16, F1 Score: 94.26±0.28	Context aware Loss (CAL) addresses the inherent class imbalance in ECG signals by dynamically adjusting the weights based on class representation
Islam et al. (2024)	MIT-BIH (Moody and Mark 2001) + INCART (Tihonenko et al. 2007)	CAT-Net	F1 Score: 99.14%	SMOTE-Tomek was used to address class imbalance in the ECG data
Busia et al. (2024)	MIT-BIH (Moody and Mark 2001)	Tiny Transformer	Accuracy: 98.97%	The transformer model, with only 6,000 parameters, can classify the five most common arrhythmia classes
Liu et al. (2024)	Custom CHD Dataset + MIT-BIH (Moody and Mark 2001) + MIT-BIH ST Change (Albrecht 1983)	PSC-Net	Specificity: 89.93%, Sensitivity: 84.06%	A LSTM based network (GRWA-LSTM) is utilized for local temporal feature extraction
Din et al. (2024)	MIT-BIH (Moody and Mark 2001)	Hybrid Transformer Model (HTM)	F1 Score: 99.34%	Fuses features extracted by a CNN, LSTM, and transformer model to leverage the unique advantages of each
(Chon et al. 2023)	MIT-BIH (Moody and Mark 2001)	Hybrid Transformer Model (HTM)	Specificity:0.898, Sensitivity: 0.807	Multi-kernel ResNet was used to extract fetures, and postion embedding based on the heart rate was used
Hu et al. (2021)	MIT-BIH (Moody and Mark 2001)	Custom Transformer	F1 Score: 99.83%	An adaptive heartbeat segmentation method was implemented to electively focus on the time-dependent representation of heartbeats

The table compares models based on dataset usage, core methodologies, performance metrics, and unique innovations in handling raw ECG data, focusing on how different architectures leverage transformers to improve classification accuracy and efficiency

talizes on the spatial feature extraction capability of CNNs, the temporal sequence learning ability of LSTMs, and the long-range dependency capturing strength of transformers. The transformer component of the model uses 6 MHSA modules. Feature fusion layer merge outputs from CNN, LSTM, and transformer branches, and classification is done using a majority voting system. One limitation of this approach is its reliance on three separate branches, which makes it computationally intensive. Lastly, Chon et al. (2023) integrate a multi-kernel resNet (MK-ResNet) with a transformer model. The MK-ResNet employs two different kernel sizes (5 and 11) to capture features at multiple scales. These feature maps are then fed into a standard transformer architecture. The HRV-based position encoding within the transformer integrates heart rate variability data directly into the model. This structure allows the transformer to recalibrate features extracted from MK-ResNet, leading to improved classification performance. However, the model's heavy reliance on pretraining with external datasets, like icentia11k, can lead to performance degradation when applied to ECG signals from new, diverse patient populations not represented in the pretraining data.

Several studies have advanced transformer-based ECG beat classification by integrating additional preprocessing techniques and multimodal data to enhance feature extraction and address challenges like data imbalance and signal variability (refer to Table 2). Zhu et al. (2024) employ ICMT-Net, focusing on both ECG and wrist pulse signals. The model integrates an improved ConvNeXt with a multimodal transformer and MLP fusion layers, enabling the extraction and fusion of features from both signal modalities. The standard transformer architecture is modified with batch normalization layers and a convolutional block attention module. However, ICMT-Net is dependent on a carefully tuned multimodal fusion strategy, which may not generalize well to other types of physiological signals. Xia

Table 2 Summary of transformer-based models with specialized preprocessing of ECG signals and multimodal data integration for ECG beat classification

Reference (Year)	Dataset information	Core methodology	Result	Remarks
Zhu et al. (2024)	MIT-BIH (Moody and Mark 2001)	ICMT-Net	Specificity: 98.29, Sensitivity: 96.30	Uses both ECG and wrist pulse signal (WPS) for a multimodal classifier
Xia et al. (2023)	MIT-BIH (Moody and Mark 2001)	TCGAN	Accuracy: 94.69%	Generates heartbeat signals per disease type which are then added to the original dataset to address the data-imbalance problem
Xia et al. (2023)	MIT-BIH (Moody and Mark 2001)	Hybrid Transformer Model (HTM)	Specificity:99.37, Sensitivity: 99.91	Local features are extracted by a denoising encoder before being passed to a lightweight transformer
Guan et al. (2021)	MIT-BIH (Moody and Mark 2001)	LDTF	Specificity: 98.39%, Sensitivity: 98.41%	The low-dimensional denoising embedding (LDE) captures low dimensional representation of the signal in the time-frequency domain, retaining the signals temporal information

The table compares models based on dataset usage, core methodologies, performance metrics, and innovations in preprocessing and data integration techniques et al. (2023) presents a transformer and convolution-based generative adversarial network (TCGAN) for addressing the class imbalance problem in abnormal beat classification. The transformer encoder is modified by incorporating two-stage upsampling to enhance the resolution of generated ECG signals. The discriminator remains a conventional CNN, focusing on local feature extraction to differentiate between real and synthetic data. These modifications allow the model to generate high-quality ECG signals that closely resemble real data. However, the reliance on fixed-length input sequences may reduce the model's generalizability to ECG signals of varying lengths. Xia et al. (2023) introduce a novel seq2seq model that combines a lightweight transformer with CNN and denoising autoencoder (DAE) embeddings to improve inter-patient ECG arrhythmia classification. The CNN and DAE modules extract ECG wave features from individual heartbeats that are merged with manually extracted R-R features. Combined features with positional encoding are passed through a transformer encoder for capturing global dependencies. This approach leads to improved performance, particularly in minority classes. However, a limitation of the proposed model is its reliance on the pre-training of the DAE using external datasets, which introduces a dependency on data not inherent to the target domain. Lastly, Guan et al. (2021) implement a low-dimensional denoising embedding transformer (LDTF) for abnormal beat classification. The low-dimensional denoising embedding (LDE) stage combines features from both time and frequency domains using discrete wavelet transform (DWT) and fast Fourier transform (FFT). This LDE is integrated with a transformer architecture, which includes 8 standard transformer encoder layers. The blocks are strategically placed to maximize the extraction of both time-domain and frequency-domain features, enhancing the model's ability to classify arrhythmias. However, a shortcoming of this approach is its dependence on hand-crafted feature extraction techniques like DWT and FFT.

A few works have introduced innovations in attention mechanisms within transformers, developing modified or novel strategies to improve the accuracy of ECG beat classification (refer to Table 3). Meng et al. (2022) present a lightweight fussing transformer model where standard self-attention is replaced with a lightConv attention (LCA) mechanism to reduce the model's complexity while maintaining performance. The architecture incorporates a CNN-based input embedding structure, which extracts intra-heartbeat features using local attention mechanisms. The LCA reduces parameters by 72% compared to self-attention, improving efficiency without hampering accuracy. However, a critical limitation of this approach is its reliance on handcrafted convolutional structures. Tao et al. (2024) introduce a refined transformer model arrhythmia beat detection. The model uses two specialized attention mechanisms: refined diagonal attention and refined gated linear (GAL) attention. These mechanisms reduce the computational burden by selectively focusing on critical correlations between heartbeats, thus maintaining the essential temporal and morphological information needed for arrhythmia detection. The attention blocks are placed within a collaborative framework that processes both rhythmic and beat arrhythmia. This placement also accelerates model convergence by approximately 50% compared to other models. However, the model's reliance on predefined rhythm and heartbeat segmentation may restrict its ability to generalize to unsegmented ECG datasets. Wu et al. (2024) present the SRT model, a CNN-Transformer with a dilated stem module, and the spatial and channel residual gated attention (SC-RGA) modules for classifying 2D heartbeat images. The dilated stem module is placed at the input stage to expand the receptive field without losing spatial resolution. The SC-RGA module is placed after the MHSA layer to enhance the focus on critical spatial

Table 3 Summary of transformer-based architectures employing novel attention mechanisms for ECG beat classification

Reference (Year)	Dataset information	Core methodology	Result	Remarks
Tao et al. (2024)	MIT-BIH (Moody and Mark 2001)	Refined-attention transformer model	Specificity: 95.22%, Sensitivity: 97.50%	Two transformer models share rhythm information through a "collaborative block" that facilitates interaction between them
Wu et al. (2024)	MIT-BIH (Moody and Mark 2001)	Hybrid Transformer Model (HTM)	Accuracy: 95.7%, Sensitivity: 0.881, F1 Score: 0.826	Model's performance was improved by expanding the transformer's receptive field and extracting features from multiple subspaces using the SC-RGA attention mechanism
Wang et al. (2023) (2023)	MIT-BIH (Moody and Mark 2001)	ACA-MA Transformer	Accuracy: 96.61%, Specificity: 99.31%, Sensitivity: 93.27%	A linear projection layer is used to capture the semantic features of ECG signals by aligning ECG tags with their corresponding segmented signals
Meng et al. (2022) (2022)	Custom Dataset (10 dynamic single lead ECG recordings, collected using a unified wearable ECG device, sampling frequency: 400Hz, duration: 24 h)	Lightweight Fussing Transformer	Specificity: 0.9975, Sensitivity: 0.9984	LightConv Attention (LCA) is a replacement for traditional self-attention in transformers, offering comparable or superior performance while utilizing fewer parameters

The table compares various models in terms of dataset usage, core methodologies, performance metrics, and unique innovations in attention mechanisms

and channel-wise features. However, a critical limitation of this approach is its reliance on the spatial configuration of 2D ECG images. Lastly, Wang et al. (2023) propose a novel MHSA, termed ACA-MA for transformer models for ECG beat classification. ACA-MA utilizes a linear projection layer to extract semantic features from ECG signals. Moreover, a position encoding-based spatiotemporal characterization method is used to integrate time series information into a matrix format, and a MHSA to capture global contextual information. However, a critical limitation of this model is its dependency on carefully engineered position encoding, which may not adapt well to variations in ECG signals.

3.1.2 Rhythm classification

Transformer-based architectures have been popular in identifying irregular cardiac rhythms from ECG signals (refer to Table 4). Wang et al. (2021) introduce a heartbeat-aware transformer (HAT) that incorporates heartbeat position attention within the transformer block. The architecture features a convolutional backbone that processes ECG to extract ECG features. Subsequently, the model employs heartbeat-aware transformer blocks to incorporate heartbeat positions into the standard self-attention process. However, the model is limited to short ECG inputs (8 s), potentially missing long-term temporal patterns essential for accurate arrhythmia classification. Similarly, Yang et al. (2022) propose a component-aware transformer (CAT) that first decomposes ECG waveforms into components like the P-wave, QRS-complex, and T-wave using a 1D U-net-based segmentation model. These components are then converted into tokens that include information about their length and type, allowing the model to capture detailed characteristics of each waveform segment. This approach enhances the model's ability to analyze both single-lead and multi-lead ECG signals. However, the model heavily depends on the accuracy of the ECG segmentation, which was trained on a small dataset (LUDB(Kalyakulina et al. 2020) with only 200 cases), potentially limiting its generalizability.

Several works attempt to enhance traditional transformer architectures with residual and convolutional techniques. Pratiher et al. (2022) present an enhanced vision transformer (ViT) by introducing dilated convolutional and residual connections. The ECG signals are transformed into time-frequency representations (TFRs) using continuous wavelet transform (CWT), short-time fourier transform (STFT), and chirplet transform (CT). These TFRs are then stacked and used as input to the ViT for AF detection. The dilated convolutional stem expands the receptive field for multi-scale feature extraction without loss of resolution. Subsequently, eight transformer layers, with an MHSA mechanism are employed. Residual connections are integrated throughout to improve training stability and aid convergence. However, the model's reliance on time-frequency representations (TFRs) may limit its adaptability to raw ECG data, and its performance across different arrhythmia types remains underexplored. Che et al. (2021) propose a transformer architecture that is embedded within a CNN framework. Specifically, the architecture includes 8 standard transformer blocks with MHSA. Performance is enhanced by using a novel link constraint loss function that clusters embedding vectors of the same class closer together, improving the model's ability to differentiate between classes and mitigating the effects of data imbalance. This link constraint improves feature consistency and classification accuracy, but it adds significant computational costs due to the complexity, making it impractical for resource-constrained environments like portable ECG devices.

Lastly, some studies have focused on leveraging self-supervised learning and multi-scale attention mechanisms. Zou et al. (2023) employ a CNN-Transformer architecture for AF detection, with a core novelty centered around a self-supervised learning approach. The architecture employs a standard transformer consisting of three blocks with MHSA, following a four-block CNN backbone for initial feature extraction. The self-supervised learning approach involves a "next clip" prediction task, where the model is pre-trained to predict the next segment of an ECG signal based on the previous interval, enabling it to learn temporal features critical for AF detection. However, like Wang et al. (2021), the model is limited to short ECG inputs (10 s) potentially missing long-term temporal patterns. In another work, Ji et al. (2024) present a transformer architecture that combines multi-scale grid attention with self-attention mechanisms to effectively capture both spatial and temporal features of ECG signals. The model first divides the 12-lead ECG into limb and chest leads, subsequently using self-attention to fuse features from these lead subsets. Next, the multi-scale grid attention mechanism dynamically adjust grid sizes to extract temporal features at various scales, enabling the model to capture both local and global patterns in the ECG data. However, this integration of multi-scale grid and self-attention mechanisms introduces significant computational complexity, which may reduce the model's practicality for computationally-constrained environments.

Table 4 Overview of recent advancements for rhythm classification using ECG signals

Reference (Year)	Dataset	Core methodology	Result	Remarks
Ji et al. (2024)	CPSC 2018 (Alday et al. 2020) + MIT-BIH (Moody and Mark 2001)	Multi-Scale Grid Transformer (MSGformer)	F1 Score: 0.86, Sensitivity: 97.13%, Specificity: 97.87%, Accuracy: 99.28%	Spatial features captured from limb and chest leads are used by a multi-scale grid attention mechanismto capture temporal features
(Zou et al. 2023)	CPSC 2018 (Alday et al. 2020)	Self supervised learning witha CNNTransformer architecture	Sensitivity: 0.84 ± 0.01, Specificity: 0.84 ± 0.01	The model only accepted 10-second ECG recordings and showed lowered classification upon pre-training
Pratiher et al. (2022)	PCC 2017 (Clifford et al. 2017) + PCC 2021 (Reyna et al. 2021)	Dilated Residual ViT (DiResViT)	Sensitivity: 98.05%, Specificity: 98.16%	The standard patchify stem of the ViT is replaced with dilated convolutional stem with residual connections for improved detection
Yang et al. (2022)	Shaoxing Hospital ZhejiangUniversity School of Medicine database (Zheng et al. 2022)	Component-Aware Transformer (CAT)	F1 Score: 81.53−85.13%, AUC: 0.9767−0.9823	The ECG waveform is decomposed into components, which are vectorized into a single vector with length and type information, and used as the transformer's input
Wang et al. (2021)	Custom Dataset (multi-lead ECGrecords of over 200,000 patients from200+ hospitals. (sampling freq 250 Hz, leads = 4)	Heartbeat-Aware Transformer (HAT)	F1 Score: 0.9628	Introduces Heartbeat Position Attention which heartbeat positions into encoder-decoder attention
Che et al. (2021)	CPSC 2018 (Alday et al. 2020)	Transformer merged CNN architecture	F1 Score: 0.786	The quality of feature embeddings was enhanced using link constraints that enforced similar data items to have closely aligned features (Must-link) and different items to have distinctly separate features (No-link)

3.2 Sleep apnea

Several studies have focused on employing transformers to improve the detection of obstructive sleep apnea (OSA) from ECG signals (refer to Table 5). Hu et al. (2022) present a hybrid transformer model (HTM) for OSA detection using single-lead ECG signals. The model introduces a multiperspective channel attention (MPCA) block to automatically derive and fuse features from raw ECG sequences, R-peak amplitude, interbeat (RR) interval, and RR interval first-order difference. The MPCA block employs three parallel two-layer convolutional networks with varying kernel sizes to capture multiperspective features. Additionally, a squeeze-and-excitation (SE) block is integrated to adaptively focus on the most relevant features. These features along with positional encodings are fed into transformer blocks to capture long-range dependency in features. While this architecture reduces complexity by using automatically derived features, the selected features might oversimplify the ECG's complex, nonlinear characteristics, potentially limiting the model's ability to capture subtle OSA-related patterns. Similarily, Wei et al. (2024) introduce the MSAF-Transformer, a multi-scale attentive feature network for OSA prediction using single-lead ECG signals. First, the network extracts features from raw ECG signals with CNNs of varying kernel sizes. Next, the extracted features pass through an attention channel compression module that performs point-to-point convolution to reduce the number of feature channels. This decreases the overall network parameters and computational complexity. Additionally, SE blocks are employed to mitigate the impact of compression and enhance feature expression, while the transformer's self-attention mechanism captures long-distance dependencies to improve OSA detection. However, the model's reliance on pre-defined multi-scale convolutional kernels may not fully adapt to the varying temporal dynamics and feature scales present in ECG signals. Liu et al. (2023) propose a CNN-Transformer architecture for OSA detection using single-lead ECG signals. The architecture uses 10 CNN-BN-ReLU blocks for dimensional transformation, followed by 2 transformer encoder blocks designed to capture temporal dependencies through self-attention. The encoder blocks employ MHSA, focusing on global feature extraction. This hybrid approach allows for robust feature learning and classification, yet the reliance on a fixed 3-minute input window may limit adaptability to varying apnea event durations, potentially reducing detection accuracy for events longer than this window. Hu et al. (2023) present a HTM-based personalized transfer learning approach for OSA detection. The HTM architecture captures multiscale features with four parallel convolution branches and utilizes channel attention through SE modules. These features are then processed by two standard transformer blocks to capture temporal dependencies and enhance sleep apnea detection. The model also introduces a label mapping length (LML) selection strategy, which involves experimenting with different time windows for mapping signal segments to enhance the model's focus on relevant features during training. Additionally, the fine-tuning strategy is employed to adjust the model's parameters to individual patient data, ensuring better personalization of the model's predictions. However, the reliance on SE modules may lead to overemphasizing certain signal features, potentially missing subtle indicators critical for accurate apnea detection. Lastly, Li et al. (2023) propose a time-frequency information fusion-based CNN-Transformer model (TFFormer) for OSA detection using single-lead ECG signals. Unlike standard transformer architectures, the TFFormer introduces a series of specialized modules, including a deep residual shrinkage network (DRSN) for noise reduction, a multiscale convolutional atten-

Table 5 Overview of recent advancements in Sleep Apnea (SA) detection using ECG signals

Reference (Year)	Dataset	Core methodology	Result	Remarks
Wang et al. (2024)	Apnea-ECG(Penzel et al. 2000) + Private Database (PSG recordings of 49 males and 13 femailes, sampling freq: 256 Hz)	ResT-ECGAN	Sensitivity: 0.957, Specificity: 0.917	Addresses the lack of data, and low data quality by augmenting the data with a GAN network
Wei et al. (2024)	Apnea-ECG(Penzel et al. 2000)	MSAF-Transformer	Accuracy: 88.9%, Sensitivity: 0.844, Specificity: 0.917	Proposes a mutil-scale attentive feature network to extract rich local and temporal features from single-lead ECG signals
Hu et al. (2023)	Apnea-ECG (Penzel et al. 2000) + UCDDB (Heneghan 2011)	Hybrid Transformer Model (HTM)	Accuracy: 85.4%, AUC: 0.915	Model performance improved when the dataset was balanced using a random cropping strategy
Li et al. (2023)	Apnea-ECG (Penzel et al. 2000) + XJ Dataset (Shi et al. 2023)	TFFormer	Accuracy: 91.68%, Sensitivity: 89.15%, Specificity: 93.25%, F1 Score: 89.11%	Propose a gated self-attention mechanism for time-frequency information and ECG data fusion
Fayyaz et al. (2023)	NHC Sleep Data Bank (Lee et al. 2022) + CHAT (Redline et al. 2011)	Custom Transformer	F1 Score: 83.9, ROC-AUC: 90.6	The model demonstrated the highest apnea classification performance when using the combination of ECG and SpO2 signals, outperforming all other combinations of the six available PSG signals
Liu et al. (2023)	Apnea-ECG (Penzel et al. 2000)	CNNTransformer	Accuracy: 88.2%, Sensitivity: 78.5%, specificity: 94.1 %, F1 Score: 89.0, AUC: 0.947	Provides a promising and reliable solution for home portable detection of OSA
Hu et al. (2022)	Apnea-ECG (Penzel et al. 2000)	Hybrid Transformer Model (HTM)	Accuracy: 91%, ROC-AUC: 0.96, Specificity: 93.34%, Sensitivity: 86.46%, F1 Score: 87.47, MAE: 2.71, MCC: 79.86	The model inputs are four typical feature sequences are directly derived from the raw ECG data without relying on manual expert feature extraction

tion (MSCA) module for rich feature extraction, and an adaptive pruning time-frequency fusion attention (APTFFA) module. The time-frequency information fusion is achieved by separately extracting time-domain and frequency-domain features using self-attention mechanisms and then merging these through a gated mechanism that balances the contribution of each domain. These modifications enhance the model's ability to remove redundant tokens, effectively combining time- and frequency-domain information for more accurate OSA detection. However, a critical limitation of this approach is its dependency on the fixed scaling of time- and frequency-domain features, which may restrict the model's adaptability to variations in signal characteristics across different patient populations.

Some studies have focused on data augmentation techniques and data-fusion based architectures to improve model performance. Fayyaz et al. (2023) present a customized transformer-based architecture for detecting OSA, utilizing a novel data representation technique to effectively handle polysomnography (PSG) modalities. The model segments sleep signals and electronic health records (EHRs) into equal-length segments, which are then synchronized and tokenized. The tokenizer re-samples evenly spaced time series (like ECG and SpO2) and applies interpolation to irregular time series (like R-R intervals from ECG). This processed data is converted into tokens that are fed into a standard transformer encoder block for further analysis. The architecture utilizes five standard transformer encoder blocks for the detection of OSA. However, the model's reliance on PSG-derived signals may reduce its effectiveness in at-home testing scenarios where signal quality is typically lower. Wang et al. (2024) propose a ResT-ECGAN framework for OSA detection, introducing a novel combination of a transformer and ResNet, along with a GAN-based data augmentation technique (ECGAN). ECGAN is trained on combined architectures of DCGAN and LSTMs and it filters the generated ECG signals by incorporating the concept of fuzziness, effectively increasing the amount of high-quality data. The core architecture, ResT-Net (Li et al. 2018), uses a modified ResNet backbone, where standard 3x3 convolutions are replaced with 1D convolutions to reduce computational complexity, followed by a transformer encoder layer to capture dependencies between local and global features. The integration of the ECGAN for data augmentation boosts the model's performance by generating high-quality synthetic ECG segments, addressing the data scarcity issue. However, the reliance on synthetic data may introduce subtle artifacts that could impact the generalization of the model to real-world data.

3.3 Myocardial infarction

Transformer-based models have emerged as powerful tools in the detection of myocardial infarction (MI) using ECG signals (refer to Table 6). Shan et al. (2022) presents a hybrid network designed for MI localization using 12-lead ECG signals, combining the strengths of convolutional and transformer architectures. The core innovation lies in integrating lightweight depthwise separable convolutions, which reduce computational load, with an enhanced transformer that uses relative position representations. This enhancement overcomes the traditional transformer's limitation of losing accuracy when dealing with shifted ECG waveforms, thereby making the self-attention mechanism more robust for ECG analysis. The transformer block, positioned after the CNN module, fine-tunes the extracted features by focusing on the most critical regions, while the branch attention mechanism prioritizes the most informative ECG leads, improving the overall localization accuracy.

However, the model's separate processing of each ECG lead may miss crucial inter-lead relationships and dependencies, limiting its adaptability and accuracy in real-world scenarios with varied lead placements or signal abnormalities. Wahid et al. (2024) implements a hybrid model combining ResNet and ViT architectures to improve MI detection by leveraging both global and local features. The core novelty is found in the modification of the standard ViT, where a slim model incorporates multibranch networks and a channel attention mechanism. This modification involved replacing the conventional 16×16 convolution for patch embedding with a series of smaller convolutions, enhancing the richness of the extracted features. The hybrid architecture strategically places these blocks to optimize the integration of ResNet's local features with ViT's global features, leading to a more comprehensive feature representation. However, the reliance on standard dense layers for feature alignment between ResNet and ViT can potentially lead to suboptimal feature integration due to the inherent differences in the feature extraction processes of the two models.

ResNet-based architectures continue to play a crucial role in ECG classification, particularly in enhancing model performance and convergence. Zhang et al. (2023) introduce MSFNet, a model that leverages a multi-stage architecture, combining ResNet-18 and transformer Encoder blocks to improve MI classification from ECG signals. The architecture improves upon standard ResNet by introducing weight sharing, accelerating convergence, and boosting performance. A streamlined 6-layer transformer block refines feature extraction while minimizing parameters and training time. The Position Attention Module (PAM) highlights and localizes key ECG segments by aggregating spatial features and assigning higher attention weights to abnormal areas, thereby improving classification accuracy. However, the potential loss of subtle inter-lead relationships in ECG signals due to the architecture's reliance on separate processing branches, may lead to less accurate MI localization in complex cases.

Table 6 Overview of recent advancements in myocardial infarction (MI) detection using ECG signals

Reference (Year)	Dataset information	Core methodology	Result	Remarks
Liu et al. (2024)	PTB (Bousseljot et al. 1995)	SRTNet	Sensitivity: 99.15%, Specificity: 98.58%	ECG diagnosis by integrating scanning, reading, and thinking modules to mimic a doctor's analysis
Wahid et al. (2024)	Custom Dataset (1500 cases, of different types ofMI)	Hybrid Transformer Model (HTM)	Specificity: 0.96, Sensitivity: 0.94	Fuses ResNet and ViT feature vectors to create a robust representative feature vector
Zhang et al. (2023)	PTB-XL (Wagner et al. 2020)	MSFNet	Sensitivity: 52.14%, Specificity: 77.52%, F1 Score: 53.89%	Uses a modified ResNet architecture with weights sharing to speed up model convergence
Shan et al. (2022)	PTB (Bousseljot et al. 1995)	Mobi-Trans	Specificity: 99.91%, Sensitivity: 99.91%	Designed a 12-lead ECG signal acquisitiondevice with Wi-Fi Transmission for remote monitoring
Liu et al. (2022)	MITNSR (Goldberger et al. 2000) + BIDMC (Pimentel et al. 2016)	ECVT-Net	Specificity: 99.96%, Sensitivity: 99.96%	Inter-patient and intra-patient evaluations were conducted, along with anti-noise testing using ECGs of varying quality to assess model robustness

Custom network architectures offer specialized solutions for MI detection, enhancing the diagnostic capabilities of ECG-based systems. Liu et al. (2024) presents a novel architecture named SRTNet, designed for MI detection and localization using 12-lead ECG signals. The core novelty lies in the integration of three specialized modules-Scanning, Reading, and Thinking-each addressing different aspects of ECG analysis. The architecture significantly modifies standard transformer elements by incorporating large-kernel convolutions in the Scanning module and grouped convolutions in the Reading module, enhancing feature extraction and reducing parameter count. The Thinking module uses a simplified transformer to extract temporal features, strategically positioned after spatial feature extraction to capture comprehensive ECG information. However, a limitation of this approach is the underperformance in MI localization tasks, likely due to the insufficient and imbalanced data used for training, which hinders the model's ability to generalize effectively across different MI types. Liu et al. (2022) introduces ECVT-Net, a novel architecture that integrates CNNs with ViTs for detecting congestive heart failure (CHF) using ECG signals. The core novelty of this work is the combination of the local feature extraction capabilities of CNNs with the global dependency modeling of ViTs. Unlike standard ViT models, the model has a transition block that splits the CNN-derived feature maps into fixed-size sequences and applies a linear transformation before feeding them into the ViT. This modification enhances the model's ability to capture both short-term and long-term dependencies in the ECG signals. Standard transformer blocks are strategically placed after the convolutional layers to optimize the integration of local and global features. However, the paper lacks investigation into the interpretability of the model, particularly in understanding how the transformer layers contribute to decision-making, which is essential for clinical applications.

3.4 Miscellaneous diagnosis and analysis

Several studies have addressed the task of leveraging transformers for comprehensive CVD diagnosis (refer to Table 7). Vaid et al. (2023) introduce HeartBEiT, a ViT-based designed to enhance the diagnostic performance of ECGs, with the application being the detection of cardiac conditions such as hypertrophic cardiomyopathy (HCM), low left ventricular ejection fraction (EF), and ST-elevation MI. The core novelty lies in its use of masked image modeling (MIM) for pre-training on ECG data, which allows the model to learn from vast amounts of unlabeled ECGs by treating ECG waveforms as 2D images. The standard ViT architecture is modified by converting ECG signals into 14x14 patches, tokenized using the Dall-E model. These modifications significantly improve diagnostic accuracy, particularly in limited-data scenarios. However, the model relies on accurate patch tokenization; any inaccuracies in this process could impair the model's ability to detect subtle diagnostic features. Similarly, Fu et al. (2024) propose CardioGPT, a novel approach for ECG interpretation that integrates a wavelet scattering network having a transformer-based architecture, designed to generate natural language interpretations of ECGs. The core novelty lies in adapting the GPT architecture, traditionally used for text, to interpret 12-lead ECG signals by treating them as a sequence of vectors. The architecture uses standard transformer blocks specifically adapted to handle the 3D input variables generated by the wavelet scattering transformation. A limitation of the model is its dependency on the accuracy of wavelet scattering for feature extraction, which may affect its performance in cases of noise or signal distortion.

Table 7 Overview of recent transformer-based methodologies for detection of various cardiovascular conditions using ECG signals

Reference (Year)	Application	Dataset	Core methodology	Result	Remarks
Fu et al. (2024)	Natural language interpretations of ECG	Private Dataset (1,128,553 ECG readings from 754,920 patients, 10 sec, 12 channel, sampling rate: 500 Hz)	CardioGPT	BLEU: 0.68, ROUGE: 0.78	The model builds on the ContrastiveLanguage -Image Pretraining (CLIP) model developed by OpenAI
Vaid et al. (2023)	Cardiac condition diagnosis	MSHS ECG dataset (private) (Vaid et al. 2023)	HeartBEiT	MI Detection AUROC:0.94, HCM AUROC: 0.92, EF: 0.93	The model was pre-trained on over 8.5 million ECGs showing capablities to detect various cardiac diseases
Wang et al. (2023)	Coronary Artery Disease Detection (CAD)	PTB (Bousseljot et al. 1995) + PKUSZ Diagnostic CAD Database (private) (Wang et al. 2023)	MF-CADNet	Sensitivity: 97.74%, Specificity: 95.83%	The mode incorporates information from ECG, Spectrogram, Vectrocardiogram (VCG) and electronic medical records (EMR) to make accurate detection of CAD
Ryu et al. (2023)	Left Ventricular Hypertrophy (LVH)	Private Dataset (34,302 subjects [19,044 male 15,258 female subjects], over 31 days, 12-lead, 10sec long ECG, sampling rate:500 Hz)	CoAt-Mixer	Mean sensitivity 78.37%, Mean Specificity 74.04%	Constructed a learning environment wherein the gender differences were leveraged to better classify LVH
Behinaein et al. (2021)	Stress Detection	WESAD (Schmidt et al. 2018) + SWELL-KW (Koldijk et al. 2014)	Hybrid Transformer Model (HTM)	Accuracy: 91.1%, F1 Score: 83.3	The model is an end-to-end network comprising three subnetworks, a convolutional subnetwork, a transformer encoder, and a fully connected (FC) subnetwork

Recent advancements in transformer-based models have shown significant promise in the detection of CVDs by effectively leveraging multi-modal medical data. Wang et al. (2023) present MF-CADNet, a transformer-based multi-feature fusion network designed for the detection of coronary artery disease (CAD) using a combination of ECG, vectrocardiogram (VCG), spectrograms, and electronic medical records (EMR). The core novelty of the work lies in its integration of multiple data modalities with SeResNet modules handling initial feature extraction and a transformer encoder performing inter-module information fusion.

The transformer architecture in MF-CADNet uses standard MHSA, but these are strategically placed to combine and extract high-dimensional features from the various input modalities, enhancing the model's diagnostic accuracy. However, the model is dependent on the accuracy of the VCG generation from ECG; as any inaccuracies in this conversion process could negatively impact the overall diagnostic performance. Similarily, Ryu et al. (2023) presents CoAt-Mixer, a self-attention-based framework designed for the detection of left ventricle hypertrophy (LVH) using ECG. The core novelty of the work lies in utilizing a hybrid model that integrates MBConv blocks with a CBAM attention module for feature extraction, followed by transformer blocks for capturing long-range dependencies. The modifications, such as replacing the SE module with the CBAM module, help improve the attention mechanism within the convolutional process, enhancing the model's ability to focus on relevant ECG features. A limitation of this approach is the potential oversimplification in handling complex ECG variations, as the model heavily relies on predefined convolutional operations, which might not fully capture subtle yet clinically significant variations in ECG signals. Behinaein et al. (2021) employs a transformer-based architecture for stress detection using ECG signals. The convolutional blocks are specifically tasked with extracting spatio-temporal features, which are then enhanced by positional encoding before being fed into the transformer. The architecture uses standard transformer blocks, including MHSA. However, the model's dependency on fine-tuning with subject-specific data to achieve competitive performance may limit its scalability in real-world applications where individual calibration is not feasible.

3.5 LLMs for ECG analysis

LLMs have been gaining traction for ECG analysis and diagnosis due to their sophisticated understanding of natural language, enabling them to merge knowledge from multiple sources and expertly interpret ECGs (refer to Table 8). Yu et al. (2023) proposed a zero-shot retrieval-augmented diagnosis technique for diagnosing arrhythmia and sleep apnea. The first stage involves the construction of a vector database of ECG diagnosis books, which is used to retrieve relevant diagnostic information based on extracted ECG features (i.e., diagnostic guidance). The diagnostic information is combined with feature prompts and fed into LLMs (i.e., LLaMA2 and GPT−3.5) to generate a structured diagnostic output and explanations. One key limitation of this work is that the diagnostic capability of the LLMs is constrained by the information derived from ECG textbooks and the handcrafted ECG features available in open-source databases. Oh et al. (2024) present ECG-QA for training and fine-tuning LLMs for ECG analysis. The dataset includes 70 distinct question-answering templates that can be categorized into two main types: 1) query a single ECG and 2) comparative analysis between two different ECGs. The single ECG questions encompass a wide range of attributes, including SCP codes (e.g., identifying first-degree AV block), noise types (e.g., baseline drift or static noise), stages of infarction, detection of extrasystoles, heart axis deviations, and numeric features such as RR interval deviations. The comparative questions, on the other hand, focus on differences between two ECGs, such as resolving symptoms, comparing RR intervals, and other diagnostic changes, facilitating a deeper and more comprehensive ECG analysis. Xu et al. (2024) explore the idea of LLMs interacting with IoT sensors and actuators, termed "Penetrative AI". The acquired ECG data is downsampled and quantized before being fed to LLM (i.e., ChatGPT), which is guided by a fixed

Table 8 Overview of LLM-based methodologies for ECG-based diagnosis and report generation

Reference (Year)	Application	Dataset	Core methodology	Result	Remarks
Oh et al. (2024)	ECG question answering dataset for LLMs	ECG-QA (Oh et al. 2023)	Diverse collection of questions relevant for ECG analysis	Test exact match accuracy: single-verify: 76% single-choose: 58.2% single-query: 40.0	ECG-QA includes questions for comparative analysis of two different ECGs
Xu et al. (2024)	Explore LLMs ability to interact with IoT sensor data	MIT-BIH (Moody and Mark 2001)	Utilize LLMs for analyzing down sampled &	MAE (beats/min): ChatGPT-4 with expert knowledge: 1.92	LLMs are capable of analyzing ECG waveform changes (e.g., heartbeat) from raw downsampled ECG signals
Yu et al. (2024)	Arrhythmia and ECG-based subject detection	CSN (Zheng et al. 2020) MIMIC-IV-ECG (Gow et al. 2023) PTB-XL (Wagner et al. 2020)	Fuse clinically enhanced ECG description with ECG waveform feature for ECG analysis and diagnosis	Arrhythmia AUC: 93.9 User-based identification AUC: 0.97 (PTB-XL)	Clinically enhanced description of ECG signal improve understanding of deep learning models enabling better diagnosis than standard self-supervised methodologies
Wan et al. (2024)	ECG report generation with LLMs and multimodal instructions	MIMIC-IV-ECG (Gow et al. 2023) PTB-XL (Wagner et al. 2020)	Multimodal ECG Instruction Tuning (MEIT) framework	BLEU-1 performance: Mistral-Instruc: 0.714 (MIMIC-IV-ECG) LLaMA-2: 0.515 (PTB-XL)	Instruction tuning significantly improves performance across all LLMs and metrics
Liu et al. (2024)	ECG Self-supervised Learning (eSSL)	MIMIC-ECG (Gow et al. 2023) PTB-XL (Wagner et al. 2020) CPSC2018 (Liu et al. 2018) CSN (Zheng et al. 2020)	Multi-modal ECG Representation Learning (MERL) framework with Clinical Knowledge Enhanced Prompt Engineering (CKEPE)	Linear Probing (100%) results: PTB-XL Super: 88.67 PTB-XL Sub: 84.72 PTB-XL Form: 79.65 PTB-XL Rhythm: 88.34 CPSC2018: 90.57 CSN: 87.95	Enhancing prompts with clinical knowledge is crucial for MERL outperforming other eSSL approaches
Chen et al. (2024)	Heart failure risk prediction	UK-HYP (Sudlow et al. 2015) UK-MI (Sudlow et al. 2015)	ECG dual attention network (ECG-DAN)	C-index scores: UKB-HYP: 0.6349, UKB-MI: 0.5805	LLM-informed pre-training is curcial for risk prediction
Yu et al. (2023)	Zero-shot arrhythmia and sleep apnea diagnosis	PTB-XL(Strodthoff et al. 2023) + Apnea-ECG-Penzel et al. (2000)	LLM-based ECG diagnosis supported by ECG retrieval-augmented QA	Accuracy: 75.7 Macro Sensitivity: 79.1 Macro Specificity: 61.6 Macro F1: 66.9	Proposed zero-shot approach outperforms few-shot approches and achieves competitive performance to supervised learning methods

prompt to identify R-peaks. The prompt encodes expert knowledge and directs ChatGPT to evaluate the ECG sequence, tasking it to identify significant upward trends in ECG and select the highest points as R-peaks. This method illustrates the potential of LLMs to handle real-time ECG signal processing tasks, such as ECG feature detection and segmentation, within IoT frameworks. One key shortcoming of this work is that it is confined to detecting basic ECG features and does not extend to identifying more complex patterns, such as arrhythmias. Yu et al. (2024) learn enhanced ECG representations using a multimodal contrastive pertaining framework. Two core components of this framework include cardio query assistant (CQA) and ECG semantic intergator (ESI). First, CQA constructs a vector database of ECG-specific knowledge. When CQA is queried with specific ECG conditions is utilizes LLMs to retrieve relevant information, mapping ECG conditions to enriched waveform patterns. Next, the ESI component integrates enhanced ECG explanation (i.e., textual prompts) with corresponding ECG signals using a dual-encoder architecture that separately processes text and ECG signals. subsequently, a cross-modality decoder is employed captioning and contrastive losses to align and unify the textual and waveform representations, enabling the model to capture nuanced clinical insights essential for arrhythmia diagnosis and ECG-based user identification tasks. One shortcoming of this approach is that it solely relies on 10-second ECG signals without accounting for temporal duration variability in the real world. Wan et al. (2024) propose a multimodal ECG instruction tuning (MEIT) framework for generating ECG reports using LLMs. Instruction tuning is performed on LLMs with autoregressive objectives, leading to report ECG-based report generation. The MEIT architecture includes an ECG encoder that processes multi-lead ECG data and integrates it with a text-based report generation model, optimized using LoRA (Low-Rank Adaptation). Experiments with nine different LLM backbones underscore MEIT's superior performance, robustness in zero-shot scenarios, and resilience to signal perturbations.

Liu et al. (2024) introduce a multimodal ECG representation learning (MERL) framework utilizing ECG waveform as well as the report. MERL framework employs Cross-Modal Alignment (CMA) to align representations between the ECG signals and corresponding clinical reports (generated using different encoders), utilizing cross-modal contrastive losses. The framework also incorporates Uni-Modal Alignment (UMA), which applies latent augmentation through independent dropout operations on ECG embeddings to prevent pattern corruption. The authors also propose clinical knowledge-enhanced prompt engineering (CKEPE) that generates enhanced prompts by linking LLMs with expert-verified clinical knowledge databases. Results show that MERL outperforms other ECG self-supervised learning approaches in zero-shot classification and linear probing tasks. The main limitation of this approach is that the use of LLMs for extracting ECG expert knowledge is not a fully controlled and transparent process. Chen et al. (2024) propose the use of an LLM-informed neural network for enhanced heart failure prediction. The proposed ECG-DAN (ECG Dual Attention Network) simultaneously captures cross-lead relationships and temporal dynamics within each lead, thereby enhancing feature aggregation. During pretraining, LLM is employed to generate text embeddings for structured reports (i.e., SCP code-based reports), which are aggregated based on their confidence scores. These aggregated embeddings are aligned with ECG embeddings using an alignment loss to ensure that the textual and ECG representations are closely related. During fine-tuning, encoder pre-trained weights are utilized to generate ECG embeddings, which are passed through a fully connected network for risk prediction. One shortcoming of this work is that it discounts the use of metadata (e.g., age, sex, noise presence) for generating textual embeddings.

3.6 Overarching commonalities and limitations

This subsection synthesizes the commonalities and limitations of transformer-based and LLM-based methodologies for ECG-driven diagnosis, focusing on their architectural designs and roles within the diagnostic workflow. Transformer-based approaches are primarily dictated by the type and preprocessing of ECG data, leading to four principal categories: (1) representation learning from raw ECG signals using convolutional layers followed by MHSA blocks (Hu et al. 2022; Wei et al. 2024). (2) segmentation of raw ECG signals into sub-waveforms, also processed through MHSA blocks (Hu et al. 2021; Yang et al. 2022). (3) transformation of raw ECG signals into time-frequency representations, such as CWT, STFT, or CT, followed by ViTs (Pratiher et al. 2022). (4) conversion of raw ECG signals into images, subsequently processed by ViTs (Wu et al. 2024). In parallel, LLM-based frameworks for ECG diagnosis are distinguished by the function LLMs serve in the diagnostic pipeline. These methods can be categorized as: (1) constructing expert knowledge databases from CVD books and articles and linking them with LLMs, enabling zero-shot diagnoses by integrating ECG features with textual knowledge (Yu et al. 2023). (2) exploiting extensive contextual understanding and long-range dependency capabilities of LLMs to directly analyze downsampled raw ECG data for feature extraction (Xu et al. 2024). (3) facilitating multimodal learning by aligning textual explanations of ECGs generated by LLMs with ECG features derived from CNNs or transformers (Yu et al. 2024). Figure 4 provides a visual representation of these methodologies, highlighting the distinct yet interconnected approaches of transformer and LLM-based techniques in advancing ECG-driven diagnostics.

A common trend in the literature is to innovate within the network stem or transformer encoder to address the unique challenges in ECG data. Several studies enhance the feature representation before passing them to standard transformer blocks by integrating core modules from other architectures in the network stem, such as GRUs (Liu et al. 2024), LSTM networks (Din et al. 2024), and ResNet (Wahid et al. 2024), to refine the feature representation of ECG signals processed by the transformer encoder. Additionally, some studies introduce novel loss functions, such as link constraint loss (Che et al. 2021), to enhance the differentiation between disease classes. On the other hand, some studies have modified the self-attention mechanism and the transformer encoder block to enhance the ECG diagnosis performance. To elaborate, Li et al. (2023) have proposed the TFFormer, which performs time and frequency information fusion by independently extracting the time and frequency domain features using MHSA, combining them using a gated mechanism accounting for contribution from each domain. Wang et al. (2021) have introduced a heartbeat-aware transformer which incorporated the heartbeat positions using r-peaks to improve the efficacy of the standard self-attention mechanism. Zhang et al. (2023) have developed MSFNet which incorporates the position attention module responsible for localizing key ECG segments and assigning higher attention weight to abnormal areas.

A prominent trend in recent literature is the employment of GANs to address class imbalance and to develop lightweight, efficient architectures for ECG analysis. Specifically, convolution-based GANs have been proposed for generating ECG abnormal beats in underrepresented classes, demonstrating their efficacy in augmenting minority class data (Xia

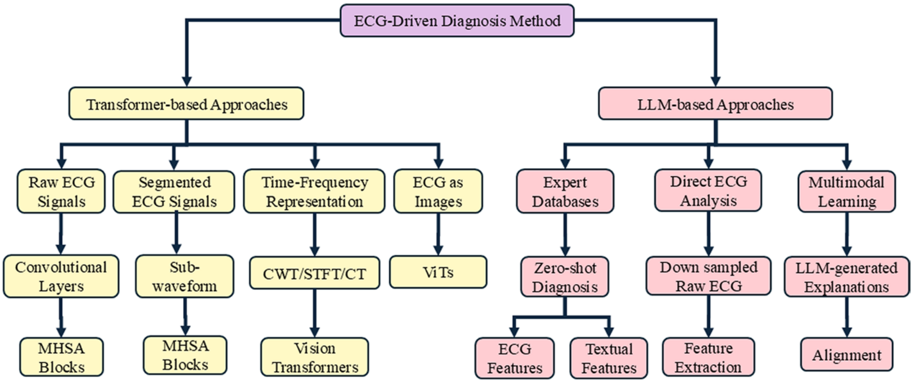

Fig. 4 Overview of ECG-driven diagnosis methods using transformer-based and LLM-based approaches

et al. 2023). Moreover, the integration of fuzzy logic with one-dimensional GANs has been explored to produce high-quality ECG data for OSA detection (Wang et al. 2024). Concurrently, researchers have focused on optimizing transformer-based methodologies to reduce computational overhead in ECG-driven diagnostics. For instance, Meng et al. (2022) have proposed the use of efficient local attention for ECG embedding generation and global light-conv attention. Busia et al. (2024) have tuned the number of MHSA blocks, ECG embedding size, and number of heads in their architecture to make it compatible with low-power devices with power consumption as low as 0.09 mJ per inference. Shan et al. (2022) have focused on minimizing the parameters in the network stem used for generating ECG embedding by using depthwise separable convolutions with SE blocks inspired by MobileNet-V3. Collectively, these advancements in network stem design, along with the refinement of ECG embedding sizes and MHSA blocks, have significantly advanced transformer-based ECG methods, facilitating their deployment on wearable and IoT devices.

Despite the advancements in transformer-based methods for ECG diagnosis, several limitations persist, specifically in the areas of ECG embeddings, positional encodings, and the design of the MHSA block. Several methodologies rely on representation learning through convolutional layers to generate ECG embeddings. However, whether these embeddings effectively distinguish between different ECG classes within the embedding space remains unclear. This ambiguity may cause a performance bottleneck as a poorly defined embedding space can severely limit the ability of MHSA blocks to optimize performance across diagnostic classes. Furthermore, most current approaches employ standard sinusoidal positional encodings originally designed for the NLP tasks. While the standard encoding is effective in capturing the sequence of positions in linear data, it may not be well-suited for the specific requirements of ECG signals. ECG signals are characterized by repetitive structures, such as the P-wave, QRS complex, and T-wave, across successive heartbeats. A more tailored positional encoding can help the model to link similar structures across different beats (e.g., P-waves) while still recognizing the temporal order of beats. Without such an encoding, the model may struggle to effectively leverage the recurring patterns in ECG waveforms, potentially impairing its ability to accurately capture the temporal and morphological details essential for diagnosis. Additionally, many existing methodologies often utilize standard MHSA blocks, developed primarily for NLP applications. This approach may not fully align with the unique properties of ECG signals, such as the slight variations in repetitive beat waveforms, potentially leading to overfitting and limiting the model's ability to accurately interpret waveform duration and morphology. Moreover, ECG diagnosis benefits from a multi-perspective analysis, where information from multiple leads is integrated to form a comprehensive understanding of cardiac electrical activity. However, the standard MHSA lacks a structured, dedicated approach to handle these specific requirements, further limiting its effectiveness in ECG signal processing.

LLM-based methodologies for ECG analysis face significant limitations related to generalizability, reliance on predefined data sources, and adaptability to real-world clinical variability. One key issue is the over-reliance on existing ECG literature (e.g., textbooks and academic articles) for crafting expert databases without accounting for bias and errors, raising concerns over the correctness of the final predicted diagnosis/outcome. Another essential concern of LLM-based methods is the lack of transparency and fact-checking while extracting expert knowledge from ECG vector databases, raising doubts about the reliability and interpretability of the clinical insights generated. Furthermore, many LLM-based approaches ignore critical metadata, such as patient demographics and noise presence, during the construction of textual embeddings. Given the genetic underpinning of ECG signals and their variation with age and gender, this could lead to significant oversights in the model's predictions across different demographics and populations. Collectively, these limitations underscore the need for more adaptive, transparent, and context-aware approaches that can effectively accommodate the inherent complexity and variability of ECG data in clinical settings.

4 Discussion

This section discusses data and architectural enhancements needed to improve the state-of-the-art transformer and LLM techniques in ECG diagnosis and analysis. Furthermore, it provides potential directions to overcome existing shortcomings of LLMs.

4.1 Data enhancements and architectural innovations

4.1.1 Self-supervised and unsupervised ECG representation learning

Recently, novel approaches have been proposed in various domains for enhancing the quality of data embeddings that are subsequently fed to the MHSA blocks within the transformer encoder. Zerveas et al. (2021) propose an unsupervised pretraining setup for enhancing embedding for multi-variate time series representation learning. To elaborate, a proportion r of each variable sequence within the input is independently masked. The model's objective during this pretraining phase is to predict the complete input vectors from the masked input sequences. The predictions on the masked values are then evaluated using a mean squared error (MSE) loss, which guides the optimization of the model's parameters to enhance the fidelity of the learned representations. This pretraining strategy is crucial for developing robust embeddings, as it forces the model to capture essential temporal dependencies and feature interactions within the data, despite the presence of noise and missing information. Another work by Li et al. (2021) also capitalizes on the pertaining for enhancing the embedding quality for ViTs. Specifically, the authors first demonstrate that sparse self-attention hampers the ability of the model to capture fine-grained dependencies among the image patches. To compensate for this shortcoming, the authors propose a region-matching pretraining task that forces the model to accommodate the fine-grained dependencies of the patches within the embeddings, thereby improving the quality of learned representations. This approach can be adopted for transformer-based methods working with images of ECG beats to learn dependencies between the different sub-waveforms (e.g., p-wave and QRS complex). Wu et al. (2024) propose the use of self-supervised learning using a set of signal transformations as the pretext task to learn robust, modality-invariant representation of signals. Specifically, permutation, time-warping, noise addition, magnitude-warping, and cropping transformations are used to enhance the signal representations/embeddings. In the context of ECG signals, signal transformation pertaining tasks can be used as an effective alternative to other data augmentation techniques (e.g., GANs) to learn distinctive ECG embeddings corresponding to minority classes. Eldele et al. (2021) introduce temporal contextual contrastive learning for enhanced representation of time series signals. Specifically, in temporal contrastive learning, a strongly augmented version of the data is used to predict future timesteps of a weakly augmented version, and vice versa. Subsequently, in contextual contrastive learning, the model maximizes the similarity between matching contexts from the two different augmented views and minimizes the similarity with contexts derived from other samples in the batch. Consequently, this framework can be employed to generate ECG embeddings as an alternative to the CNN-based network stem. Zhang et al. (2023) propose a representation learning framework using a cross-reconstruction transformer. First, the time series data is transformed into the frequency domain to calculate magnitude and phase components. These data components are then segmented into patches and some patches are randomly dropped. The generated embeddings from time, magnitude, and phase modalities are passed to a cross-domain transformer encoder, which integrates information across the domains. During pretraining, the cross-domain representations are passed to a transformer decoder for reconstructing the original time series, magnitude, and phase data, facilitating robust representation learning. For downstream tasks, the model condenses the cross-domain embeddings into compact representations, optimizing them for further applications such as classification.

Altogether, these unsupervised and self-supervised pretraining strategies ranging from masked input reconstruction, region matching, signal transformations, temporal and contextual contrastive learning, and cross-domain reconstruction can provide sophisticated and enriched ECG embeddings suitable for MHSA blocks.

4.1.2 Enhanced positional encoding

ECG signals exhibit recurring morphology across successive heartbeats. Implementing a specialized positional encoding can enhance the model's ability to associate/link analogous structures (e.g., P-waves) across different heartbeats while preserving the temporal sequence of beats. The conventional sinusoidal positional encoding lacks flexibility, as it is manually parameterized and lacks learnable parameters. Furthermore, this positional encoding strategy limits the maximum length of the input sequence. Liu et al. (2020) propose FLOATER, which encodes positional information via a continuous dynamical model in a data-driven and parameter-efficient manner. Consequently, the continuous dynamic model allows the transformer to handle variable length inputs. This approach can allow ECG-based transformers to work with varying sampling rate signals, enhancing their applicability in healthcare settings. Ke et al. (2020) demonstrate that absolute positional encoding (e.g., sinusoidal encoding) when added to data embeddings brings mixed correlations due to two different heterogeneous sources. Consequently, the authors propose a transformer with untied positional encoding (TUPE) that separately computes word contextual correlation and positional correlation with different parameterizations and then adds them together. Furthermore, TUPE makes the ’CLS’ token positionless in the relative positional encoding matrix, enabling the capture of global information. Subsequently, the modifications made in TUPE can be employed for ECG positional encoding to capture relative positions of ECG sub-waveforms (e.g., p-wave) across beats and minimize mixed correlation between positional and ECG information. recently Su et al. (2024) proposed a rotary position embedding (RoPE), which encodes the relative position of tokens by multiplying the context representations with a rotation matrix. RoPE enables the decay of inter-token dependency with increasing relative distances and enables relative positional encoding for linear self-attention. Given these properties of RoPE, it can be employed for ECG-based transformers to better capture the context within ECG sub-waveforms, such as the QRS complex. However, the nature of decay in RoPE may need to be adjusted depending on the CVD diagnosis task. For instance, rhythm-based abnormalities span over multiple cardiac beats and may require a slow decay to effectively capture information about the underlying arrhythmia. In contrast, beat-based abnormalities can adopt a modified RoPE with a faster decay of inter-token dependency.

Overall, these innovations in positional encodings, involving learnable encodings (Kenton et al. 2019; Lan et al. 2019; Clark et al. 2020), separate position parametrization (Ke et al. 2020), relative position encoding (Dai et al. 2019; Huang et al. 2020; Raffel et al. 2020), and rotary transformations (Su et al. 2024) could play a significant role in enhancing transformer-based ECG diagnosis methodologies.

4.1.3 Custom self-attention mechanisms

Standard self-attention computes the weights of each sequence element with every other element. This calculation can be computationally expensive for long sequences (e.g., high sampling rate ECG or long duration ECG) and may also lead to overfitting due to repetitive structure ECG waveform. Consequently, custom sparse self-attention (Child et al. 2019) can be employed to minimize computations and resolve overfitting. Guo et al. (2019) introduce the Starformer, a model inspired by star-shaped weight computation among sequence elements. Specifically, Starformer emphasizes weights between adjacent elements while discounting relationships between distant elements, a mechanism known as band attention. Additionally, it computes attention weights between a central, influential element within the sequence (the star node) and all other elements, referred to as global attention. This self-attention topology is particularly relevant for beat abnormality analysis in ECG data, with potential for further refinement. For instance, the neighborhood defined by band attention could be adapted to focus on specific waveform segments, such as the p-wave. Furthermore, the R-peak could be designated as the star node, enabling the model to capture relationships between ECG sub-waveforms and the central R-peak. This approach could potentially enhance beat abnormality detection while reducing computational complexity. Beltagy et al. (2020) present the Longformer, an extension of the Starformer that generalizes its approach by incorporating band attention for capturing local dependencies and allowing for multiple global attention elements within a sequence. This architecture holds potential for the efficient analysis of multi-beat ECG signals, where the R-peak of each beat could serve as a global attention element, facilitating the detection of rhythmic abnormalities, such as arrhythmias. Moreover, the Longformer can be further enhanced by replacing band attention with block attention, which computes the weights between elements within a defined patch or block (e.g., an ECG beat). Subsequently, the relationships between these blocks can be captured using global attention, with each beat potentially having multiple global attention nodes (e.g., the P-wave peak, R-peak, or T-wave peak). Another crucial modification of self-attention that could improve the performance of transformers for ECG analysis is self-attention priors. Guo et al. (2019) demonstrate that dependencies in text can be more effectively captured when the importance of neighboring tokens is emphasized while the influence of distant tokens is diminished. To achieve this, they propose the Gaussian transformer, which realigns self-attention scores using a modified Gaussian prior. This approach can be similarly applied to ECG waveforms, where Gaussian priors can focus on specific waveforms, such as the P-wave, to compute wave-focused features. Subsequent encoder layers can then integrate these wave-specific features, enhancing the accuracy of ECG diagnosis. A mathematical formulation of integration of priors is as follows:

$$\text{Scores} = \text{softmax}\left(\alpha \cdot \frac{QK^\top}{\sqrt{d_k}} + (1-\alpha) \cdot P\right) V \tag{5}$$

Here, P represents the prior matrix and α is the weight controlling the contribution of the prior.

To enhance the computational efficiency of transformer-based models for real-time ECG monitoring, optimizations can be introduced within the self-attention mechanism and in the overall architectural design. Within the self-attention block, computational cost can be significantly reduced by employing sparse attention (Child et al. 2019), approximate attention (Wang et al. 2020), or attention-free (Poli et al. 2023) mechanisms. Sparse attention (Child et al. 2019) selectively attends to a subset of tokens rather than all tokens, reducing the computational complexity. Approximate attention (Wang et al. 2020) further reduces the computational burden by using low-rank factorization or kernel-based methods to approximate the attention matrix, maintaining performance while improving efficiency. Attention-free (Poli et al. 2023) mechanisms, such as gated convolutions or linearized attention, eliminate the quadratic complexity of traditional self-attention, making real-time inference more feasible. Beyond self-attention, the overall transformer architecture can be optimized by replacing the standard feed-forward network (FFN) with a mixture-of-experts (MoE) approach (Fedus et al. 2022), where a learned router directs inputs to specialized smaller FFNs, activating only a subset of parameters per inference step. In addition, early exit mechanisms can be introduced at multiple depths (Xin et al. 2020), allowing confident predictions to bypass deeper layers, reducing inference latency while maintaining accuracy. Together, these modifications improve the real-time applicability of transformer-based ECG analysis in clinical settings.

4.2 Overcoming shortcomings of LLMs in ECG Diagnosis

Two critical challenges hindering the widespread adoption of LLMs in healthcare settings are the issue of hallucination and the absence of reliable metrics to assess the confidence of LLM-generated responses. Hallucination in the context of LLMs refers to the generation of outputs that are factually inaccurate, fabricated, or not grounded in the input data or real-world context. This phenomenon can significantly hinder diagnostic accuracy, as hallucinated information can mislead medical professionals, especially when integrated into clinical decision support systems. Particularly, hallucinations are concerning in high-stakes environments pertaining to diagnosis, treatment planning, and patient outcomes. These erroneous generations can also introduce biases into downstream tasks, potentially amplifying risks in automated healthcare workflows. Dziri et al. (2022) demonstrate that lack of relevant factual data and repeated data are the main causes of LLM hallucination. Hernandez et al. (2022) show that the performance of large LLMs decreases to their counterparts with nearly half the parameters by introducing 10% redundant data in the training set. Recently, Zhou et al. (2024) compare the relative importance of unsupervised pretraining versus instruction tuning and reinforcement learning. Their findings indicate that the majority of knowledge in LLMs is acquired during the pretraining phase, with only minimal instruction tuning data required to guide models in generating high-quality outputs. Given the findings of Dziri et al. (2022) and Hernandez et al. (2022) on hallucination, as well as the importance of pretraining highlighted by Zhou et al. (2024), it can be concluded that a large-scale ECG corpus should be developed from scratch. Specifically, this corpus should comprehensively cover the underlying physics of ECG signals, the interpretation of ECG waveforms and leads, clinically relevant ECG features and their thresholds for age groups, and the ECG diagnosis knowledge base of expert electrophysiologists, while minimizing data redundancies to reduce the risk of hallucinations.

LLMs tend to be overconfident in their responses and can be influenced to reflect the demands of users through repeated prompting. This can be a significant drawback in clinical systems as medical practitioners do not have established techniques to quantify the confidence of LLM's response. *Uncertainty quantification (UQ)* is an emerging research area that aims to understand the reliability and limitations of LLMs (Zhao et al. 2024). Xiong et al. (2024) propose a consistency-based uncertainty estimation approach, which involves inducing randomness during the answer generation phase or introducing misleading hints in the prompt to assess whether the LLM's response varies. If the model's response remains consistent despite these variations, it can be inferred that the model is more confident in its output. Alternatively, Duan et al. (2024) combine the token level uncertainties in the predicted response to quantify the overall uncertainty in the LLMs. Specifically, the authors propose shifting attention to relevant (SAR) that focuses on the most meaningful tokens and sentence components, rather than treating all tokens as equally important, thereby providing more accurate uncertainty assessments. These UQ approaches have to be integrated into LLM methodologies for ECG analysis and diagnosis to enhance transparency. Subsequently, UQ could lay the foundation for integrating LLM-driven ECG systems in clinics and hospitals.

4.3 Federated learning for resource-constrained environments, scalability, and patient privacy

Federated learning (FL) is a decentralized learning framework that enables collaborative model training across multiple devices without requiring raw data exchange, making it particularly well-suited for resource-constrained healthcare settings. Conventionally, neural networks and LLMs require centralized data aggregation, which is often infeasible in clinical environments due to strict patient privacy regulations and limited infrastructure in under-resourced hospitals. In the context of ECG-based disease diagnosis, FL facilitates the development of robust, generalizable models by leveraging heterogeneous data distributions across diverse patient populations while ensuring that sensitive ECG signals remain confined to local hospital systems. Consequently, FL can be a key learning paradigm for deploying LLMs for ECG-based diagnosis in resource-constrained environments. Xu et al. (2023) propose an FL-based framework that integrates differential privacy (DP) to ensure formal privacy guarantees. A fundamental challenge in applying DP to FL is that the required noise magnitude increases with model size, often leading to degraded convergence. To mitigate this issue, the authors introduce Partial Embedding Updates (PEU), a strategy that selectively updates only a subset of model embeddings, thereby reducing the payload size and limiting the impact of DP noise on training stability. Furthermore, to address the computational constraints, the methodology incorporates Low-Rank Adaptation (LoRA) to optimize parameter efficiency and Noise Contrast Estimation (NCE) to minimize memory overhead. While federated learning ensures privacy-preserving training, it does not inherently reduce model size, leading to high inference latency, excessive memory demands, and communication overhead. To address these limitations, Yao et al. (2025) propose FedSpine, a framework that integrates PEFT with a structured pruning approach to enable the deployment of LLMs in resource-constrained environments. FedSpine employs an iterative process of pruning and fine-tuning, progressively reducing model complexity while maintaining task performance. Additionally, to account for the heterogeneous computing and communication capacities of different edge devices, the framework leverages a multi-armed bandit (MAB) algorithm to adaptively determine optimal pruning ratios and low-rank adaptation (LoRA) ranks for each device.

FL supports scalability in ECG-based diagnosis by enabling hospitals, as decentralized nodes, to incrementally contribute to model improvements without requiring a centralized data repository. As participating hospitals increase, the system scales efficiently by leveraging parallel updates from multiple institutions. Beyond just node scalability, federated learning also supports data scalability, allowing models to generalize across hospitals with diverse patient populations and ECG acquisition conditions. Computational scalability is maintained by optimizing model pruning and adaptation per node, ensuring that even low-resource hospitals and edge devices can participate without excessive computational overhead. Additionally, communication scalability is preserved through asynchronous updates and hierarchical aggregation, minimizing network congestion as the system expands. These scalability mechanisms collectively enable federated learning to support large-scale, real-world deployment of LLM-based ECG diagnosis across diverse healthcare infrastructures.

In summary, FL offers a viable solution for deploying LLMs in ECG-based disease diagnosis in resource-constrained environments by preserving patient privacy across hospitals through differential privacy mechanisms and mitigating computational and memory overhead via structured pruning, iterative fine-tuning, and parameter-efficient adaptation techniques such as PEFT.

4.4 Dataset considerations for large language models in ECG analysis and diagnosis

With the advent of LLMs, new tasks related to ECG have emerged, such as ECG captioning, ECG report generation, and multimodal ECG diagnosis, which were not previously explored with standard convolutional networks or transformer-based architectures. Prior reviews have comprehensively studied existing datasets for ECG disease diagnosis using CNNs (Ansari et al. 2023). However, key challenges remain regarding dataset bias, quality, and diversity, particularly for LLM-based ECG analysis.

Current approaches in multimodal ECG learning generally fall into two categories: (1) leveraging existing metadata, such as SCP codes or machine-generated diagnostic reports, for contrastive learning to enhance ECG-text representations, and (2) retrieval-augmented generation (RAG) methods, where LLMs are supplemented with external knowledge bases. Most publicly available datasets for these tasks, including PTB-XL (predominantly representing the German population), MIMIC-IV, and Chapman-Shaoxing (primarily covering Chinese populations), lack diversity in Middle Eastern, South Asian, and African populations. This raises concerns about model generalizability, as ECG morphology is influenced by genetic and demographic factors. Additionally, while RAG-based methods enhance zero-shot ECG diagnosis by integrating external databases, the clinical reliability of these databases remains uncertain. Although they may be sourced from reputable repositories, their population coverage can introduce biases, potentially limiting the model's performance.

Another challenge pertains to preprocessing and encoding strategies for ECG metadata. Different studies have adopted varied encoders to transform textual ECG metadata into feature representations for contrastive learning. However, there is no consensus on the most effective encoding framework, leading to inconsistencies in how textual information is incorporated into multimodal models. Establishing standardized methodologies for encoding ECG metadata is critical for ensuring reproducibility and improving model robustness.

5 Conclusion

The manuscript provides a comprehensive survey of transformer-based methodologies and their evolution into LLMs for ECG analysis and diagnosis. The reviewed studies are systematically categorized according to the complexity of disease diagnosis, spanning single-beat to multi-beat scenarios. A key contribution is the detailed analysis of overarching patterns and limitations, which reveals that CNNs are predominantly utilized for ECG representation learning, GANs for data augmentation, and standard MHSA blocks as the preferred choice for capturing temporal dependencies. Notable limitations include a lack of novel ECG representation techniques, insufficient positional encoding formulations, and the absence of task-specific self-attention structures. For LLMs, limitations include limited generalizability, dependency on unverified pre-existing databases, and insufficient mechanisms for transparency and fact-checking. Based on these findings, the discussion introduces promising future directions focusing on enhanced ECG representation, refined positional encodings, custom self-attention structures, and strategies to mitigate hallucinations and improve confidence estimation in LLMs. Additionally, future work will require close collaboration between machine learning experts and clinical practitioners to ensure that LLM-derived ECG diagnosis aligns with real-world clinical workflows and decision-making processes. Effective deployment will depend on integrating domain knowledge into model training and validating LLM-driven insights with expert feedback. Furthermore, the regulatory approval process for LLM-based medical diagnosis remains a critical challenge. Establishing standardized evaluation frameworks, demonstrating model robustness across diverse populations, and addressing ethical concerns related to LLM-driven healthcare decisions will be essential for widespread adoption.

Funding Open Access funding provided by the Qatar National Library.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Gustafsson S, Gedon D, Lampa E, Ribeiro AH, Holzmann MJ, Schön TB, Sundström J (2022) Development and validation of deep learning ecg-based prediction of myocardial infarction in emergency department patients. Sci Rep 12(1):19615

van Klei WA, Bryson GL, Yang H, Kalkman CJ, Wells GA, Beattie WS (2007) The value of routine preoperative electrocardiography in predicting myocardial infarction after noncardiac surgery. Annals of surg 246(2):165–170

Milstein NS, Musat DL, Allred J, Seiler A, Pimienta J, Oliveros S, Bhatt AG, Preminger M, Sichrovsky T, Shaw RE et al (2020) Detection of atrial fibrillation using an implantable loop recorder following cryptogenic stroke: implications for post-stroke electrocardiographic monitoring. J Int Card Electrophys 57:141–147

Bouzid Z, Al-Zaiti SS, Bond R, Sejdić E (2022) Remote and wearable ecg devices with diagnostic abilities in adults: a state-of-the-science scoping review. Heart Rhythm 19(7):1192–1201

Ha AC, Verma S, Mazer CD, Quan A, Yanagawa B, Latter DA, Yau TM, Jacques F, Brown CD, Singal RK et al (2021) Effect of continuous electrocardiogram monitoring on detection of undiagnosed atrial fibrillation after hospitalization for cardiac surgery: a randomized clinical trial. JAMA Netw Open 4(8):e2 121 867-e2 121 867

Dickinson DF (2005) The normal ecg in childhood and adolescence. Heart 91(12):1626–1630

Rijnbeek PR, Van Herpen G, Bots ML, Man S, Verweij N, Hofman A, Hillege H, Numans ME, Swenne CA, Witteman JC et al (2014) Normal values of the electrocardiogram for ages 16–90 years. J Electrocard 47(6):914–921

Kumari L, Sai YP et al (2022) Classification of ecg beats using optimized decision tree and adaptive boosted optimized decision tree. Sig, Image and Video Process 16(3):695–703

Mert A, Kılıç N, Akan A (2014) Evaluation of bagging ensemble method with time-domain feature extraction for diagnosing of arrhythmia beats. Neural Compt Appl 24:317–326

Shi H, Wang H, Huang Y, Zhao L, Qin C, Liu C (2019) A hierarchical method based on weighted extreme gradient boosting in ecg heartbeat classification. Comp Methods and Prog Biomed 171:1–10

Rashed-Al-Mahfuz M, Moni MA, Lio' P, Islam SMS, Berkovsky S, Khushi M, Quinn JM (2021) Deep convolutional neural networks based ecg beats classification to diagnose cardiovascular conditions. Biomed Eng Lett 11:147–162

Makimoto H, Höckmann M, Lin T, Glöckner D, Gerguri S, Clasen L, Schmidt J, Assadi-Schmidt A, Bejinariu A, Müller P et al (2020) Performance of a convolutional neural network derived from an ecg database in recognizing myocardial infarction. Sci Rep 10(1):8445

Karthiga M, Santhi V, Sountharrajan S (2022) Hybrid optimized convolutional neural network for efficient classification of ecg signals in healthcare monitoring. Biomed Sig Process Control 76:103731

Peng H, Chang X, Yao Z, Shi D, Chen Y (2024) A deep learning framework for ecg denoising and classification. Biomed Sig Process Control 94:106441

Ji C, Wang L, Qin J, Liu L, Han Y, Wang Z (2024) Msgformer: a multi-scale grid transformer network for 12-lead ecg arrhythmia detection. Biomed Sig Process Control 87:105499

Wang Z, Pan X, Mei Z, Xu Z, Lv Y, Zhang Y, Guan C (2024) Ecgan-assisted rest-net based on fuzziness for osa detection. IEEE Trans Biomed Eng 71(8):2518–2527

Liu K, Liu T, Wen D, Zang M, Zhou S, Liu C (2024) Srtnet: Scanning, reading, and thinking network for myocardial infarction detection and localization. Expert Systems with Applications 240:122402

Qiu J, Zhu J, Liu S, Han W, Zhang J, Duan C, Rosenberg MA, Liu E, Weber D, Zhao D (2023)“Automated cardiovascular record retrieval by multimodal learning between electrocardiogram and clinical report,” in *Proceedings of the 3rd Machine Learning for Health Symposium* ser. Proceedings of Machine Learning Research, (225), 480–497

Thapa S, Howlader K, Bhattacharjee S et al. (2024) “More: Multi-modal contrastive pre-training with transformers on x-rays, ecgs, and diagnostic report,” arXiv preprint arXiv:2410.16239

Guo M, Zhou Y, Tang S (2024) “Multimodal models for comprehensive cardiac diagnostics via ecg interpretation,” in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 5756–5763

Yu H, Guo P, Sano A (2023) “Zero-shot ecg diagnosis with large language models and retrieval-augmented generation,” in *Proceedings of the 3rd Machine Learning for Health Symposium*, ser. Proceedings of Machine Learning Research, vol. 225, 10 Dec, pp. 650–663

Chua CE, Clara NLY, Furqan MS, Kit JLW, Makmur A, Tham YC, Santosa A, Ngiam KY (2024) “Integration of customised llm for discharge summary generation in real-world clinical settings: a pilot study on russell gpt,” *The Lancet Regional Health–Western Pacific*, vol. 51

Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, Demner-Fushman D, Dligach D, Daneshjou R, Fernandes C et al. (2025) “The tripod-llm reporting guideline for studies using large language models,” *Nature Medicine*, pp. 1–10

Han C, Zhou Y, Que W, Li Z, Shi L (2024) “An overview of algorithms for myocardial infarction diagnostics using ecg signals: Advances and challenges,” *IEEE Transactions on Instrumentation and Measurement*,

Maron BJ, Desai MY, Nishimura RA, Spirito P, Rakowski H, Towbin JA, Rowin EJ, Maron MS, Sherrid MV (2022) Diagnosis and evaluation of hypertrophic cardiomyopathy: jacc state-of-the-art review. J Am Coll Cardiol 79(4):372–389

Ebrahimi Z, Loni M, Daneshtalab M, Gharehbaghi A (2020) A review on deep learning methods for ecg arrhythmia classification. Exp Syst with Appl: X 7:100033

Dehkordi NR, Dehkordi NR, Toudeshki KK, Farjoo MH (2024) Artificial intelligence in diagnosis of long qt syndrome: a review of current state, challenges, and future perspectives. Mayo Clin Procee: Digit Health 2(1):21–31

Salari N, Hosseinian-Far A, Mohammadi M, Ghasemi H, Khazaie H, Daneshkhah A, Ahmadi A (2022) Detection of sleep apnea using machine learning algorithms based on ecg signals: a comprehensive systematic review. Exp Syst Appl 187:115950

Liu X, Wang H, Li Z, Qin L (2021) Deep learning in ecg diagnosis: a review. Knowl-Based Syst 227:107187

Nezamabadi K, Sardaripour N, Haghi B, Forouzanfar M (2022) Unsupervised ecg analysis: a review. IEEE Rev Biomed Eng 16:208–224

Ansari MY, Qaraqe M, Righetti R, Serpedin E, Qaraqe K (2024) Enhancing ecg-based heart age: impact of acquisition parameters and generalization strategies for varying signal morphologies and corruptions. Front Cardiovas Med 11:1424585

Wasimuddin M, Elleithy K, Abuzneid A-S, Faezipour M, Abuzaghleh O (2020)“Stages-based ecg signal analysis from traditional signal processing to machine learning approaches: A survey,” *IEEE Access*, (8), 177 782–177 803

Ansari MY, Qaraqe M, Charafeddine F, Serpedin E, Righetti R, Qaraqe K (2023) “Estimating age and gender from electrocardiogram signals: A comprehensive review of the past decade,” *Artificial Intelligence in Medicine*, p. 102690

Yokota A, Kabutoya T, Hoshide S, Kario K (2021) Automatically assessed p-wave predicts cardiac events independently of left atrial enlargement in patients with cardiovascular risks: the japan morning surge-home blood pressure study. J Clin Hypertension 23(2):301–308

Rasmussen MU, Kumarathurai P, Fabricius-Bjerre A, Larsen BS, Domínguez H, Davidsen U, Gerds TA, Kanters JK, Sajadieh A (2020) P-wave indices as predictors of atrial fibrillation. Annals of Noninvasive Electrocard 25(5):e12751

Wu S, Cai M, Zheng R, Wang S, Jiang L, Xu L, Shi R, Xiao F, Ellenbogen KA, Cha Y et al (2021) Impact of qrs morphology on response to conduction system pacing after atrioventricular junction ablation. ESC Heart Failure 8(2):1195–1203

Pelliccia A, Tatangelo M, Borrazzo C, Zampaglione D, Mango F, Fedele E, Lanzillo C, Martino A, Crescenzi C, Maestrini V et al (2023) Low qrs voltages and left ventricular hypertrophy: a risky association. Eur J Prev Cardiol 30(11):1132–1138

Luo G, Li Q, Duan J, Peng Y, Zhang Z (2020) The predictive value of fragmented qrs for cardiovascular events in acute myocardial infarction: a systematic review and meta-analysis. Front Phys 11:1027

Li R, Zhao X, Gong Y, Zhang J, Dong R, Xia L (2021) A new method for detecting myocardial ischemia based on ecg t-wave area curve (twac). Front Phys 12:660232

D'Ascenzi F, Anselmi F, Adami PE, Pelliccia A (2020) Interpretation of t-wave inversion in physiological and pathological conditions: current state and future perspectives. Clin Cardiol 43(8):827–833

Gillioz A, Casas J, Mugellini E, Abou Khaled O (2020) "Overview of the transformer-based models for nlp tasks," in *2020 15th Conference on computer science and information systems (FedCSIS)*. IEEE, pp. 179–183

Tetko IV, Karpov P, Van Deursen R, Godin G (2020) State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. Nat Commun 11(1):5575

Yaqoob M, Ishaq M, Ansari MY, Konagandla VRS, Tamimi TA, Tavani S, Corradetti A, Seers TD (2024) Geocrack: a high-resolution dataset for segmentation of fracture edges in geological outcrops. Sci Data 11(1):1–13

Yaqoob M, Ishaq M, Ansari MY, Qaiser Y, Hussain R, Rabbani HS, Garwood RJ, Seers TD (2025) Advancing paleontology: a survey on deep learning methodologies in fossil image analysis. Artif Intell Rev 58(3):83

Yaqoob M, Ansari MY, Ishaq M, Jayachandran ISAJ, Hashim M, Seers TD (2025) "Microcrystalnet: An efficient and explainable convolutional neural network for microcrystal classification using scanning electron microscope petrography," *IEEE Access*

Dahmani H, Yaqoob M, Ansari MY, Flushing EF (2025) "Thermal homography in photovoltaic panels: Evaluating deep learning and feature-based methods," in, IEEE Texas Power and Energy Conference (TPEC). IEEE 2025:1–6

Yaqoob M, Ansari MY, Ishaq M, Ashraf U, Pavuluri S, Rabbani A, Rabbani HS, Seers TD (2025) "Fluidnet-lite: Lightweight convolutional neural network for pore-scale modeling of multiphase flow in heterogeneous porous media," *Advances in Water Resources*, p. 104952

Dong L, Xu S, Xu B (2018) "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in, IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE 2018:5884–5888

Karita S, Chen N, Hayashi T, Hori T, Inaguma H, Jiang Z, Someki M, Soplin NEY, Yamamoto R, Wang X (2019) "A comparative study on transformer vs rnn in speech applications," in, et al IEEE automatic speech recognition and understanding workshop (ASRU). IEEE 2019:449–456

Ansari MY, Yang Y, Meher PK, Dakua SP (2023) Dense-psp-unet: a neural network for fast inference liver ultrasound segmentation. Comp Biol Med 153:106478

Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H (2023) Transformers in medical imaging: a survey. Med Image Analy 88:102802

Ansari MY, Mangalote IAC, Meher PK, Aboumarzouk O, Al-Ansari A, Halabi O, Dakua SP (2024) "Advancements in deep learning for b-mode ultrasound segmentation: A comprehensive review," *IEEE Transactions on Emerging Topics in Computational Intelligence*

Li J, Chen J, Tang Y, Wang C, Landman BA, Zhou SK (2023) Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. Med Image Analy 85:102762

Ansari MY, Yang Y, Balakrishnan S, Abinahed J, Al-Ansari A, Warfa M, Almokdad O, Barah A, Omer A, Singh AV et al (2022) A lightweight neural network with multiscale feature enhancement for liver ct segmentation. Sci Rep 12(1):14153

Akhtar Y, Dakua SP, Abdalla A, Aboumarzouk OM, Ansari MY, Abinahed J, Elakkad MSM, Al-Ansari A (2021) Risk assessment of computer-aided diagnostic software for hepatic resection. IEEE Trans Radiat Plasma Med Sci 6(6):667–677

Ansari MY, Abdalla A, Ansari MY, Ansari MI, Malluhi B, Mohanty S, Mishra S, Singh SS, Abinahed J, Al-Ansari A et al (2022) Practical utility of liver segmentation methods in clinical surgeries and interventions. BMC Med Imag 22(1):97

Ansari MY, Qaraqe M (2023) Mefood: a large-scale representative benchmark of quotidian foods for the middle east. IEEE Access 11:4589–4601

Zhang M, Qiu L, Chen Y, Yang S, Zhang Z, Wang L (2023) A conv-transformer network for heart rate estimation using ballistocardiographic signals. Biomed Sig Process Control 80:104302

Lih OS, Jahmunah V, Palmer EE, Barua PD, Dogan S, Tuncer T, García S, Molinari F, Acharya UR (2023) Epilepsynet: novel automated detection of epilepsy using transformer model with eeg signals from 121 patient population. Comp Biol Med 164:107312

Afsa I, Ansari MY, Paul S, Halabi O, Alataresh E, Shah J, Hamze A, Aboumarzouk O, Al-Ansari A, Dakua SP (2024) “Development and validation of a class imbalance-resilient cardiac arrest prediction framework incorporating multiscale aggregation, ica and explainability,” *IEEE Transactions on Biomedical Engineering*

Salloum R, Kuo C-CJ (2017) “Ecg-based biometrics using recurrent neural networks,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.IEEE, pp. 2062–2066

Hou B, Yang J, Wang P, Yan R (2019) Lstm-based auto-encoder model for ecg arrhythmias classification. IEEE Trans Instrumen Measur 69(4):1232–1240

Takase S, Kobayashi S (2020) All word embeddings from one embedding. Adv Neural Inf Process Syst 33:3775–3785

Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y (2024) Roformer: enhanced transformer with rotary position embedding. Neurocomputing 568:127063

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Lu, Polosukhin I (2017) “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30

Voita E, Talbot D, Moiseev F, Sennrich R, Titov I (2019) “Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned,” arXiv preprint arXiv:1905.09418

Geva M, Schuster R, Berant J, Levy O (2020) “Transformer feed-forward layers are key-value memories,” arXiv preprint arXiv:2012.14913

El-Ghaish H, Eldele E (2024) Ecgtransform: empowering adaptive ecg arrhythmia classification framework with bidirectional transformer. Biomed Sig Process Control 89:105714

Moody GB, Mark RG (2001) The impact of the mit-bih arrhythmia database. IEEE Eng Med Biol Magazine 20(3):45–50

Bousseljot R, Kreiseler D, Schnabel A (1995) “Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet,”

Islam MR, Qaraqe M, Qaraqe K, Serpedin E (2024) Cat-net: convolution, attention, and transformer based network for single-lead ecg arrhythmia classification. Biomed Sig Process Control 93:106211

Tihonenko V, Khaustov A, Ivanov S, Rivin A et al. (2007) “St.-petersburg institute of cardiological technics 12-lead arrhythmia database,” *Dataset on physionet. org*

Busia P, Scrugli MA, Jung VJ-B, Benini L, Meloni P (2024) “A tiny transformer for low-power arrhythmia classification on microcontrollers,” *IEEE Transactions on Biomedical Circuits and Systems*

Liu Q, Feng Y, Xu H, Li J, Lin Z, Li S, Qiu S, Wu X, Ma Y, Xu Y et al (2024) Psc-net: integration of convolutional neural networks and transformers for physiological signal classification. Biomed Sig Process Control 91:106040

Albrecht P (1983) “St segment characterization for long term automated ecg analysis,” Ph.D. dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering...,

Din S, Qaraqe M, Mourad O, Qaraqe K, Serpedin E (2024) Ecg-based cardiac arrhythmias detection through ensemble learning and fusion of deep spatial-temporal and long-range dependency features. Artif Intell Med 150:102818

Chon S, Ha K-W, Park S, Jung S (2023) “An ecg beat classification method using multi-kernel resnet with transformer,” in *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, pp. 140–144

Hu S, Cai W, Gao T, Zhou J, Wang M (2021) Robust wave-feature adaptive heartbeat classification based on self-attention mechanism using a transformer model. Physiological Measurement 42(12):125001

Zhu J, Feng Y, Liu Q, Xu H, Miao Y, Lin Z, Li J, Liu H, Xu Y, Li F (2024) “An improved convnext with multimodal transformer for physiological signal classification,” *IEEE Access*

Xia Y, Xu Y, Chen P, Zhang J, Zhang Y (2023) Generative adversarial network with transformer generator for boosting ecg classification. Biomed Sig Process Control 80:104276

Xia Y, Xiong Y, Wang K (2023) A transformer model blended with cnn and denoising autoencoder for inter-patient ecg arrhythmia classification. Biomed Sig Process Control 86:105271

Guan J, Wang W, Feng P, Wang X, Wang W (2021) “Low-dimensional denoising embedding transformer for ecg classification,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1285–1289

Tao Y, Xu B, Zhang Y (2024) “Refined self-attention transformer model for ecg-based arrhythmia detection,” *IEEE Transactions on Instrumentation and Measurement*

Wu W, Huang Y, Wu X (2024) Srt: improved transformer-based model for classification of 2d heartbeat images. Biomed Sig Process Control 88:105017

Wang Y, Yang G, Li S, Li Y, He L, Liu D (2023) Arrhythmia classification algorithm based on multi-head self-attention mechanism. Biomed Sig Process Control 79:104206

Meng L, Tan W, Ma J, Wang R, Yin X, Zhang Y (2022) Enhancing dynamic ecg heartbeat classification with lightweight transformer model. Artif Intell Med 124:102236

Alday EAP, Gu A, Shah AJ, Robichaux C, Wong A-KI, Liu C, Liu F, Rad AB, Elola A, Seyedi S et al (2020) Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. Physiol Measur 41(12):124003

Zou C, Müller A, Martens E, Müller P, Rückert D, Steger A, Becker M, Wolfgang U (2023) "Self-supervised learning for atrial fibrillation detection with ecg using cnntransformer," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 807–812

Pratiher S, Srivastava A, Priyatha YB, Ghosh N, Patra A (2022) "A dilated residual vision transformer for atrial fibrillation detection from stacked time-frequency ecg representations," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.IEEE, pp. 1121–1125

Clifford GD, Liu C, Moody B, Li-wei HL, Silva I, Li Q, Johnson A, Mark RG (2017)"Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017," in *2017 Computing in Cardiology (CinC)*.IEEE, pp. 1–4

Reyna MA, Sadr N, Alday EAP, Gu A, Shah AJ, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S (2021) "Will two do? varying dimensions in electrocardiography: the physionet, computing in cardiology challenge 2021," in, et al Computing in Cardiology (CinC), vol. 48. IEEE 2021:1–4

Yang M-U, Lee D-I, Park S (2022) Automated diagnosis of atrial fibrillation using ecg component-aware transformer. Comp Biol Med 150:106115

Zheng J, Guo H, Chu H (2022) "A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0. 0)," *PhysioNet 2022Available online httpphysionet orgcontentecg arrhythmia10 0accessed on*, vol. 23,

Wang B, Liu C, Hu C, Liu X, Cao J (2021) "Arrhythmia classification with heartbeat-aware transformer," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.IEEE, pp. 1025–1029

Che C, Zhang P, Zhu M, Qu Y, Jin B (2021) Constrained transformer network for ecg signal processing and arrhythmia classification. BMC Med Inf Dec Making 21(1):184

Kalyakulina A, Yusipov I, Moskalenko V, Nikolskiy A, Kosonogov K, Zolotykh N, Ivanchenko M (2020) "Lobachevsky university electrocardiography database," *Type: Dataset. Available online: https://physionet. org/content/ludb/1.0. 0/(accessed on 10 July 2021)*

Wang Z, Pan X, Mei Z, Xu Z, Lv Y, Zhang Y, Guan C (2024) "Ecgan-assisted rest-net based on fuzziness for osa detection," *IEEE Transactions on Biomedical Engineering*

Penzel T, Moody GB, Mark RG, Goldberger AL, Peter JH (2000) "The apnea-ecg database," in Computers in Cardiology, Vol. 27 (Cat. 00CH37163). IEEE 2000:255–258

Wei C, Kuang H, Ma X, Liu X (2024) "Msaf-transformer: A multi-scale attentive feature network for sleep apnea prediction based on single lead ecg signal," in, IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), vol. 6. IEEE 2024:1009–1013

Hu S, Wang Y, Liu J, Yang C (2023) "Personalized transfer learning for single-lead ecg-based sleep apnea detection: exploring the label mapping length and transfer strategy using hybrid transformer model," *IEEE Transactions on Instrumentation and Measurement*

Heneghan C (2011)"St. vincent's university hospital/university college dublin sleep apnea database," *Vincent's university hospital/University College Dublin sleep apnea database*

Li C, Shi Z, Zhou L, Zhang Z, Wu C, Ren X, Hei X, Zhao M, Zhang Y, Liu H et al. (2023) "Tfformer: A time frequency information fusion based cnn-transformer model for osa detection with single-lead ecg," *IEEE Transactions on Instrumentation and Measurement*

Shi Y, Cao Z, Xie Y, Yuan Y, Chen X, Su Y, Niu X, Liu H, Yin L, Zhao B et al (2023) Association between obstructive sleep apnea and thyroid function: a 10-year retrospective study. Sleep Med 103:106–115

Fayyaz H, Strang A, Beheshti R (2023) "Bringing at-home pediatric sleep apnea testing closer to reality: A multi-modal transformer approach," in *Machine Learning for Healthcare Conference*. PMLR, pp. 167–185

Lee H, Li B, DeForte S, Splaingard ML, Huang Y, Chi Y, Linwood SL (2022) A large collection of real-world pediatric sleep studies. Sci Data 9(1):421

Redline S, Amin R, Beebe D, Chervin RD, Garetz SL, Giordani B, Marcus CL, Moore RH, Rosen CL, Arens R et al (2011) The childhood adenotonsillectomy trial (chat): rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population. Sleep 34(11):1509–1517

Liu H, Cui S, Zhao X, Cong F (2023) Detection of obstructive sleep apnea from single-channel ecg signals using a cnn-transformer architecture. Biomed Sig Process Control 82:104581

Hu S, Cai W, Gao T, Wang M (2022) A hybrid transformer model for obstructive sleep apnea detection based on self-attention mechanism using single-lead ecg. IEEE Trans Instrumen Measur 71:1–11

Li B, Wei W, Ferreira A, Tan S (2018) Rest-net: diverse activation modules and parallel subnets-based cnn for spatial image steganalysis. IEEE Sig Process Lett 25(5):650–654

Wahid JA, Mingliang X, Ayoub M, Husssain S, Li L, Shi L (2024) A hybrid resnet-vit approach to bridge the global and local features for myocardial infarction detection. Sci Rep 14(1):4359

Zhang L, Su X, Zheng W (2023) "Msfnet: Multi-stage fusion network for myocardial infarction classification," in *2023 4th International Conference on Computer Engineering and Intelligent Control (ICCEIC)*. IEEE, pp. 424–427

Wagner P, Strodthoff N, Bousseljot R-D, Kreiseler D, Lunze FI, Samek W, Schaeffter T (2020) Ptb-xl, a large publicly available electrocardiography dataset. Sci Data 7(1):1–15

Shan C, Zhao J, Qiu Z, Wei F, Yuan Z (2022) "Mobi-trans: A hybrid network with attention mechanism for myocardial infarction localization," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8

Liu T, Si Y, Yang W, Huang J, Yu Y, Zhang G, Zhou R (2022) Inter-patient congestive heart failure detection using ecg-convolution-vision transformer network. Sensors 22(9):3283

Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220

Pimentel MA, Johnson AE, Charlton PH, Birrenkott D, Watkinson PJ, Tarassenko L, Clifton DA (2016) Toward a robust estimation of respiratory rate from pulse oximeters. IEEE Trans Biomed Eng 64(8):1914–1923

Fu G, Zheng J, Abudayyeh I, Ani C, Rakovski C, Ehwerhemuepha L, Lu H, Guo Y, Liu S, Chu H et al. (2024) "Cardiogpt: An ecg interpretation generation model," *IEEE Access*

Vaid A, Jiang J, Sawant A, Lerakis S, Argulian E, Ahuja Y, Lampert J, Charney A, Greenspan H, Narula J et al (2023) A foundational vision transformer improves diagnostic performance for electrocardiograms. NPJ Digi Med 6(1):108

Wang X, Li J, Wang X, (2023) "Multi-feature fusion network of ecg and vcg for coronary artery disease detection," in *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things*, pp. 164–169

Ryu JS, Lee S, Chu Y, Ahn M-S, Park YJ, Yang S (2023) Coat-mixer: self-attention deep learning framework for left ventricular hypertrophy using electrocardiography. Plos one 18(6):e0286916

Behinaein B, Bhatti A, Rodenburg D, Hungler P, Etemad A (2021) "A transformer architecture for stress detection from ecg," in *Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pp. 132–134

Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K (2018)"Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM international conference on multimodal interaction*, pp. 400–408

Koldijk S, Sappelli M, Verberne S, Neerincx MA, Kraaij W (2014) "The swell knowledge work dataset for stress and user modeling research," in *Proceedings of the 16th international conference on multimodal interaction*, pp. 291–298

Oh J, Lee G, Bae S, Kwon J-m, Choi E (2024) "Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram," *Advances in Neural Information Processing Systems*, vol. 36,

Oh J, Lee G, Bae S, Kwon J-m, Choi E (2023) "Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36.Curran Associates, Inc., pp. 66 277–66 288. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/d0b67349dd16b83b2cf6167fb4e2be50-Paper-Datasets_and_Benchmarks.pdf

Xu H, Han L, Yang Q, Li M, Srivastava M (2024) "Penetrative ai: Making llms comprehend the physical world," in *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, pp. 1–7

Yu H, Guo P, Sano A (2024) "Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text," arXiv preprint arXiv:2405.19366,

Zheng J, Chu H, Struppa D, Zhang J, Yacoub SM, El-Askary H, Chang A, Ehwerhemuepha L, Abudayyeh I, Barrett A et al (2020) Optimal multi-stage arrhythmia classification approach. Sci Rep 10(1):2898

Gow B, Pollard T, Nathanson LA, Johnson A, Moody B, Fernandes C, Greenbaum N, Berkowitz S, Moukheiber D, Eslami P et al. (2023) "Mimic-iv-ecg-diagnostic electrocardiogram matched subset," *Type: dataset*

Wan Z, Liu C, Wang X, Tao C, Shen H, Peng Z, Fu J, Arcucci R, Yao H, Zhang M (2024) “Electrocardiogram instruction tuning for report generation,” arXiv preprint arXiv:2403.04945

Liu C, Wan Z, Ouyang C, Shah A, Bai W, Arcucci R (2024) “Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement,” arXiv preprint arXiv:2403.06659

Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, Liu Y, Ma C, Wei S, He Z et al (2018) An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. J Med Imaging and Health Inf 8(7):1368–1373

Chen C, Li L, Beetz M, Banerjee A, Gupta R, Grau V (2024) “Large language model-informed ecg dual attention network for heart failure risk prediction,” arXiv preprint arXiv:2403.10581

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M et al (2015) Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12(3):e1001779

Strodthoff N, Mehari T, Nagel C, Aston PJ, Sundar A, Graff C, Kanters JK, Haverkamp W, Dössel O, Loewe A et al (2023) Ptb-xl+, a comprehensive electrocardiographic feature dataset. Sci Data 10(1):279

Zerveas G, Jayaraman S, Patel D, Bhamidipaty A, Eickhoff C (2021) “A transformer-based framework for multivariate time series representation learning,” in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2114–2124

Li C, Yang J, Zhang P, Gao M, Xiao B, Dai X, Yuan L, Gao J (2021) “Efficient self-supervised vision transformers for representation learning,” arXiv preprint arXiv:2106.09785

Wu Y, Daoudi M, Amad A (2024) Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. IEEE Trans Aff Comp 15(1):157–172

Eldele E, Ragab M, Chen Z, Wu M, Kwoh CK, Li X, Guan C (2021) “Time-series representation learning via temporal and contextual contrasting,” arXiv preprint arXiv:2106.14112

Zhang W, Yang L, Geng S, Hong S (2023) “Self-supervised time series representation learning via cross reconstruction transformer,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10

Liu X, Yu H-F, Dhillon I, Hsieh C-J (2020) “Learning to encode position for transformer with continuous dynamical model,” in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119. PMLR, 13–18 Jul, pp. 6327–6335

Ke G, He D, Liu T-Y (2020) “Rethinking positional encoding in language pre-training,” in *International Conference on Learning Representations*

Kenton JDM-WC, Toutanova LK (2019) “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of naacL-HLT*, vol. 1,, p. 2

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) “Albert: A lite bert for self-supervised learning of language representations,” in *International Conference on Learning Representations*

Clark K, Luong M-T, Le QV, Manning CD (2020) “Electra: Pre-training text encoders as discriminators rather than generators,” in *International Conference on Learning Representations*

Dai Z, Yang Z, Yang Y, Carbonell JG, Le Q, Salakhutdinov R (2019) “Transformer-xl: Attentive language models beyond a fixed-length context,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988

Huang Z, Liang D, Xu P, Xiang B (2020) “Improve transformer models with better relative position embeddings,” arXiv preprint arXiv:2009.13658,

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21(140):1–67

Child R, Gray S, Radford A, Sutskever I (2019) “Generating long sequences with sparse transformers,” arXiv preprint arXiv:1904.10509

Guo Q, Qiu X, Liu P, Shao Y, Xue X, Zhang Z (2019) “Star-transformer,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Jun. pp. 1315–1325

Beltagy I, Peters ME, Cohan A (2020) “Longformer: The long-document transformer,” arXiv preprint arXiv:2004.05150

Guo M, Zhang Y, Liu T (2019) Gaussian transformer: a lightweight approach for natural language inference. Procee AAAI Conf Artif Intell 33(01):6489–6496

Wang S, Li BZ, Khabsa M, Fang H, Ma H (2020) “Linformer: Self-attention with linear complexity,” arXiv preprint arXiv:2006.04768

Poli M, Massaroli S, Nguyen E, Fu DY, Dao T, Baccus S, Bengio Y, Ermon S, Ré C (2023) “Hyena hierarchy: Towards larger convolutional language models,” in *International Conference on Machine Learning*. PMLR, pp. 28 043–28 078

Fedus W, Zoph B, Shazeer N (2022) Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. J Mach Learn Res 23(120):1–39

Xin J, Tang R, Lee J, Yu Y, Lin J (2020) “Deebert: Dynamic early exiting for accelerating bert inference,” arXiv preprint arXiv:2004.12993

Dziri N, Milton S, Yu M, Zaiane O, Reddy S (2022) “On the origin of hallucinations in conversational models: Is it the datasets or the models?” in *Annual Conference of the North American Chapter of the Association for Computational Linguistics*

Hernandez D, Brown T, Conerly T, DasSarma N, Drain D, El-Showk S, Elhage N, Hatfield-Dodds Z, Henighan T, Hume T et al., (2022) “Scaling laws and interpretability of learning from repeated data,” arXiv preprint arXiv:2205.10487

Zhou C, Liu P, Xu P, Iyer S, Sun J, Mao Y, Ma X, Efrat A, Yu P, Yu L et al. (2024) “Lima: Less is more for alignment,” *Advances in Neural Information Processing Systems*, vol. 36,

Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, Wang S, Yin D, Du M (2024)“Explainability for large language models: A survey,” *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 2, feb. [Online]. Available: https://doi.org/10.1145/3639372

Xiong M, Hu Z, Lu X, LI Y, Fu J, He J, Hooi B (2024) “Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs,” in *The Twelfth International Conference on Learning Representations*, [Online]. Available: https://openreview.net/forum?id=gjeQKFxFpZ

Duan J, Cheng H, Wang S, Zavalny A, Wang C, Xu R, Kailkhura B, Xu K (2024) “Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, Aug. pp. 5050–5063. [Online]. Available: https://aclanthology.org/2024.acl-long.276

Xu M, Song C, Tian Y, Agrawal N, Granqvist F, van Dalen R, Zhang X, Argueta A, Han S, Deng Y, Liu L, Walia A, Jin A (2023) “Training large-vocabulary neural language models by private federated learning for resource-constrained devices,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5

Yao Z, Xu Y, Xu H, Liao Y, Xie Z (2025) “Efficient deployment of large language models on resource-constrained devices,” arXiv preprint arXiv:2501.02438

Ansari Y, Mourad O, Qaraqe K, Serpedin E (2023) Deep learning for ecg arrhythmia detection and classification: an overview of progress for period 2017–2023. Front Physiol 14:1246746

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Mohammed Yusuf Ansari[1,2] · Mohammed Yaqoob[1] · Mohammed Ishaq[1] · Eduardo Feo Flushing[2] · Iffa Afsa changaai Mangalote[3] · Sarada Prasad Dakua[3] · Omar Aboumarzouk[3] · Raffaella Righetti[1] · Marwa Qaraqe[4]

✉ Mohammed Yusuf Ansari
ma1@tamu.edu

[1] Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA

[2] Carnegie Mellon University in Qatar, Doha, Qatar

[3] Department of Surgery/Clinical Advancements Department, Hamad General Hospital, Hamad Medical Corporation, Doha, Qatar

[4] College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar