Secure Retrieval-Augmented Generation in ECG Using ML based Lightweight LLMs

Sai Swathi Priya Veluri[1], Dr. Chandravathi Dittakavi[2]

[1]M. Tech Scholar, Department of Computer Science and Engineering, Gayatri, Vizag AP, India.

[2]Assosciate Professor, Department of Computer Science and Engineering, Gayatri, Vizag AP, India.

.Cite this paper as: Sai Swathi Priya Veluri, Dr. Chandravathi Dittakavi, (2025) Secure Retrieval-Augmented Generation in ECG Using ML based Lightweight LLMs, *Journal of Neonatal Surgery*, 14 (28s), 1038-1049

ABSTRACT

This project introduces a framework for appropriately adapting and adjusting machine learning (ML) techniques used to construct electrocardiogram (ECG)-based schemes. With more qualified training data given to corresponding machine learning schemes, the precision on ML-based ECG mechanisms are increased in consequence. In the proposed framework four new measure metrics are introduced to evaluate the quality of the ML training and testing data, all proposed mechanisms, metrics, and sample data with demonstrations using various ML techniques, is developed. For developing ML based ECG. The system uses retrieval-augmented generation (RAG) to provide a lightweight LLM with relevant cardiology knowledge at inference time, enabling it to diagnose cardiac conditions from ECG data without task-specific training. We extend the original methodology by deploying a small, open-source LLM (Versatile--Llama 3B) locally using the Ollama platform, ensuring patient data never leaves the premises. The proposed system aims to replicate these benefits in a secure environment. We detail the existing solutions, the proposed architecture, its advantages in privacy and cost, system requirements, and a comprehensive methodology. The outcome is an on-premise ECG diagnostic assistant that leverages both the efficiency of a small LLM and the accuracy of domain-specific retrieval, demonstrating a feasible path toward AI-assisted cardiac diagnosis without compromising data security

Keywords: *ECG Classification, GPT, GANs, LLM, Model Interpretability, Arrhythmia Detection.*

1. INTRODUCTION

Electrocardiogram (ECG) is a noninvasive mode that's used as an individual tool for cardiovascular conditions Arrhythmia Detection. ECG signal is extensively used as an abecedarian tool for the discovery and opinion of heart diseases. ECG is the record of variation of bioelectric implicit with respect to time as the mortal heart beats. It provides precious information about the functional aspects of the heart and cardiovascular system. Since ECG is the most generally recorded signal for the patient monitoring and examination process, it's important to be reliably and snappily descry the cardiac diseases. ECG can be recorded fluently with the help of face electrodes on the branches or casket. It's considered a representative signal of cardiac physiology, useful in diagnosing cardiac arrhythmia. Abnormality of the ECG shape is generally called arrhythmia. Arrhythmia is a common term for any cardiac meter that differs from normal sinus meter. Beforehand discovery of heart conditions can protract life and enhance the quality of living through applicable treatment. It's veritably delicate for croakers to dissect long ECG records in a short time duration and also the mortal eye is inadequately suited to descry the morphological changes of ECG signal continuously. From the practical point of view, for the effective diagnostics, the study of ECG pattern may have to be carried out over several hours. The volume of the data being enormous, the study is tedious and time consuming and the possibility of missing the vital information is high. A number of experimenters have reported automated bracket and discovery of twinkle patterns grounded on the features uprooted from ECG signals. utmost of them use either time or frequence sphere representation of the ECG signals as features. Depending on the features, the bracket is allowed to fete between classes. Now-a-days, the automatic ECG signal analysis faces a delicate problem due to a large variation in morphological and temporal characteristics of the ECG waveforms of different cases and the same cases. At different times, the ECG waveforms may differ for the same case to similar an extent that they're unlike each other and at the same time likewise for different types of beats. Owing to this, the beat classifiers perform well on the training data but give poor performance. Clinical experts can detect AF by analyzing ECG beat signals recorded by the Holter system. Recent years have seen an increase in the [3] use of computer-based automatic heart arrhythmia detection systems by cardiologists and physicians, highlighting the importance of automatic detection of arrhythmia in facilitating prompt and accurate treatment. The computer-based ECG signals follow the steps of pre-processing, feature extraction, and classification of abnormal signals and are detected automatically. In order to proceed with feature extraction and classification, it is necessary to first remove any existing noise components. Major sources of noise in an electrocardiogram are baseline drift, power line interference, and moving artifacts [4]. Preprocessing methods that include filtering techniques aid in the elimination of noise components, allowing for rapid processing, while various signal compression procedures reduce the computational load by favoring effective signal characteristics. ECG feature extraction [5] is the primary method for diagnosing arrhythmias through the classification of the resulting feature vectors. Feature classification is the second factor that employs various different methods for classifying ECG patterns, such as neural networks, statistical methods, machine learning algorithm classifiers, and fuzzy interference [6].

Objectives

Develop a Framework: Introduce a framework for adapting and adjusting machine learning techniques used in constructing electrocardiogram (ECG)-based schemes to improve precision.

Introduce New Metrics: Propose four new measure metrics to evaluate the quality of the ML training and testing data.

Implement Retrieval-Augmented Generation (RAG): Utilize RAG to provide a lightweight LLM with relevant cardiology knowledge at inference time, enabling it to diagnose cardiac conditions from ECG data without task-specific training.

Deploy Locally: Deploy a small, open-source LLM locally using the Ollama platform to ensure patient data never leaves the premises, enhancing privacy and security.

Evaluate Performance: Assess the performance of the proposed system using standard metrics such as accuracy, precision, recall, and F1-score.

Motivation

The motivation behind choosing this specific approach, particularly the use of lightweight LLMs and the Ollama platform, can be outlined as follows:

Privacy and Security: By deploying the LLM locally using the Ollama platform, the system ensures that sensitive patient data remains secure and confidential, addressing significant concerns related to data privacy in healthcare.

Cost-Effectiveness: Utilizing open-source components and avoiding cloud-based API fees makes the system cost-effective, which is crucial for widespread adoption in various healthcare settings.

Flexibility and Generality: The use of a lightweight LLM allows the system to perform zero-shot learning of new conditions via retrieval, making it flexible and adaptable to various diagnostic tasks without extensive retraining.

Accuracy and Efficiency: Combining the interpretive ability of an LLM with the precision of expert knowledge retrieval aims to achieve high diagnostic accuracy, comparable to specialized deep learning models, while maintaining the efficiency of a lightweight system.

Limitations

Data Quality and Quantity: The performance of the system heavily relies on the quality and quantity of the training data. Limited or biased datasets may affect the accuracy and robustness of the model.

Complex Conditions: The system might struggle with diagnosing certain complex conditions, such as nuanced electrolyte changes in ECG or combined pathologies, which may require more specialized knowledge or models.

Generalization: While the system aims to be generalizable, the effectiveness of the lightweight LLM in diverse clinical settings and with varied patient populations needs further validation.

Computational Resources: Although the system is designed to be lightweight, the computational resources required for training and inference might still be a limiting factor for some healthcare facilities with limited infrastructure.

Interpretability: While the system uses interpretability tools like SHAP and Grad-CAM, the inherent complexity of LLMs might still pose challenges in fully interpreting the decision-making process.

2. RELATED WORK

Advancing Cardiovascular Diagnostics: Integrating Machine Learning and 3D Simulation in ECG Analysis for Enhanced Arrhythmia Detection and Disease Prediction [2]

This study emphasizes the crucial role of ECG analysis in diagnosing and predicting cardiovascular disorders, the significant contribution of computational methods like machine learning in accurate heartbeat classification for arrhythmia detection and disease diagnosis, and the potential of 3D computer simulations in interpreting ECG data and generating synthetic datasets for training machine learning classifiers. The paper also emphasizes the significance of ECG analysis in diagnosing cardiovascular disorders, discusses the role of computational techniques like machine learning in advancing medical discoveries, and highlights the recent progress and growing interest in the field. The methodology in the paper involves the use of machine learning techniques and 3D computer simulations for ECG analysis, including supervised and unsupervised learning for dataset analysis, with a focus on heartbeat classification and patient diagnosis. The study also discusses the challenges of dealing with medical data for clinical applications and the role of 3D computer simulations in addressing these challenges.

Intervention effects include the classification of heartbeats, SVM specificity:99.72%, AF event identification, Sensitivity: ~97%, Specificity: ~97%, Real-time analysis using deep learning, Computer simulation models: Provided insights into the effects of diseases and treatments on the ECG.

Deciphering the Heart's Rhythm: An Explainable Deep Learning Approach for Atrial Fibrillation Detection from ECGs [3]

This research discusses the development of an explainable artificial intelligence model to detect atrial fibrillation using an electrocardiogram, emphasizing the importance of interpretability in deep learning models for accurate detection. The study's objectives were to develop an explainable deep learning model to detect atrial fibrillation using an electrocardiogram, validate its performance using diverse ECG formats, and enhance its transparency for clinical application.

The main findings in this paper were the successful detection of AF using diverse ECGs with an explainable DLM, enhanced transparency of the DLM for clinical application, and the outperformance of previous models in detecting AF The methodology involved a retrospective study using the Sejong ECG dataset for internal validation, development of two feature modules for DLM decisions, and external validation using datasets from PTB-XL, Chapman, and PhysioNet.

The intervention effects in the study show that the DLM had high accuracy in detecting atrial fibrillation (AF) using different formats of ECG, with AUC values ranging from 0.990 to 0.999. The sensitivity of the DLM ranged from 0.982 to 0.999, and the specificity ranged from 0.970 to 0.999. The explainability of features such as rhythm irregularity and absence of P-wave also showed high AUC values, ranging from 0.961 to 0.993 and

0.983 to 0.993, respectively. These results indicate that the explainable artificial intelligence methodology used in the DLM was effective in accurately detecting AF and providing insights into the reasons for its decisions.

Precision in Prediction: Evaluating Supervised Machine Learning Algorithms for Differential Diagnosis in ECG Report [4]

The study aims to use supervised machine learning to distinguish between normal and abnormal ECG reports and predict specific heart diseases, evaluating the performance of six different algorithms. The methodology in the paper involves designing a model using supervised machine learning to find anomalies in ECG reports, applying six supervised machine learning algorithms to distinguish between normal and abnormal ECG, dividing the dataset for training and testing, using Cross Validation and Random Train-Test Split for accuracy, normalizing and scaling data, creating different datasets for specific diseases, training classifiers with different algorithms, and using sample data to predict heart diseases. The main findings of this research were Logistic Regression performed well with a score above 0.90 for all diseases, Myocardial Infarction had the highest accuracy score across all algorithms, and different algorithms were effective for different diseases.

The intervention effects in the study include: Coronary Artery Disease (CAD): Logistic Regression: 90%, Naive Bayes: 83%

Myocardial Infarction (MI):

Logistic Regression: 100%, Naive Bayes: 91%

Sinus Tachycardia (ST):

Decision Tree: 97%, Nearest Neighbor: 70%

Sinus Bradycardia (SB):

Support Vector Machine: 96%, Nearest Neighbor: 69% Right

Bundle Branch Block (RBBB):

Logistic Regression: 96%, Naive Bayes: 81%,

Overall, all the algorithms gave relatively very good results for all diseases.

Decoding Heartbeats: Comparative Analysis of ANN Models for Cardiac Arrhythmia Diagnosis from ECG Data [5]

This research proposes an ANN-based system for diagnosing cardiac arrhythmia using ECG signal data, evaluates three ANN models based on classification performance measures, and discusses the importance of accurate detection of arrhythmias for clinical purposes. The MLP model exhibited high accuracy and sensitivity in diagnosing cardiac arrhythmia, while the MNN model showed superior classification specificity. The performance of the ANN models varied based on the dataset used.This research discusses the performance of different ANN models in classifying cardiac arrhythmia cases, highlighting the strengths of Multilayer Perceptron in accuracy and sensitivity, and Modular Neural Network in specificity, with variations in performance across different datasets.

The methodology involved using Artificial Neural Network models to diagnose cardiac arrhythmia from ECG signal data, training the models with the backpropagation algorithm, evaluating performance using various measures, and cleaning the dataset by replacing missing values. The study objectives were to identify cardiac arrhythmias automatically from ECG recordings and to classify diseases into normal and abnormal classes using three different neural network models.

Enhancing Early Intervention: AI Methods in Identifying Shockable ECG Rhythms and Overcoming Database Challenges [6]

The study discusses the importance of early detection of shockable ECG rhythms, the incorporation of AI methods into CAAC systems, the challenges related to the use of deep learning methods, and the need for large databases for accurate classification. It emphasizes the significance of accurate identification of shockable ECG rhythms and the role of feature extraction in achieving high detection accuracy. The paper also highlights the limitations of small databases in achieving optimal classification performance. It also emphasizes the importance of detecting shockable ECG rhythms, the role of AI in improving accuracy, and the need for advanced systems to enhance real-time detection. It also highlights the significance of accurate ECG diagnosis in designing automated external defibrillators (AEDs) and the increasing use of AI in computeraided arrhythmia classification systems. The introduction sets the stage for the paper's focus on reviewing machine and deep learning methods for detecting shockable ECG signals.

The future research recommendations include applying state-ofthe-art methods on big data, collecting new high-quality and long-duration ECG data, using compression techniques to simplify deep learning models, employing data augmentation methods, utilizing cross-validation techniques, optimizing algorithms for feature selection, balancing data, exploring lightweight transfer learning models, and considering new loss functions.

Advancing ECG Analysis: Deep Multi-Task Learning for Enhanced Accuracy and Transferability [7]

This research introduces a deep multi-task learning scheme for ECG data analysis to improve the accuracy and transferability of models, addressing challenges faced by traditional deep learning algorithms. The proposed deep multi-task learning scheme for ECG data analysis improves accuracy by up to about 5.1% and shows potential for generalizability to other datasets.

The paper discusses the application of a deep multi-task learning scheme for ECG data analysis, emphasizing the division of tasks, dataset construction, and the design of a deep parameter-sharing network to improve accuracy by up to 5.1% using the MIT-BIH database.

The methodology involved proposing a deep multi-task learning scheme for ECG data analysis, converting the problem into a multi-task learning one, constructing multiple datasets for each task, designing a deep parameter-sharing network, and conducting experiments using the MIT-BIH database.

The hypotheses tested in the study are that the deep multi-task learning scheme for ECG data analysis can improve the accuracy of ECG data analysis by up to about 5.1% compared to traditional deep learning algorithms and that the proposed scheme requires limited efforts to fine-tune the network and can enable the trained model to be well applied to other datasets.

Revolutionizing ECG Arrhythmia Detection: A Novel Inter-Patient Paradigm with SVM Classification Superiority [8]

This study aims to design and investigate an automatic classification system for ECG arrhythmia detection using a new comprehensive ECG database inter-patient paradigm separation without performing any feature extraction. The study focuses on improving the detection of minority arrhythmical classes using machine learning techniques. The research emphasizes the importance of the proposed inter-patient paradigm separation method for ECG classification without the need for complex data preprocessing for feature engineering. The study evaluates four supervised machine learning models (SVM, KNN, Random Forest, and ensemble) for classifying Normal Beat, Left Bundle Branch Block Beat, Right Bundle Branch

Block Beat, Premature Atrial Contraction, and Premature

Ventricular Contraction using real ECG records from MIT-DB. The results show that the SVM classifier outperforms the other methods in terms of accuracy, precision, recall, and f1-score, achieving an accuracy of 0.83. The paper concludes that the SVM model is more realistic in a clinical environment for classifying various types of ECG signals collected from different patients.

3. ANALYSIS OF PROBLEM

Early detection and continuous monitoring of cardiac health can significantly improve patient outcomes. Electrocardiogram (ECG) signals are widely used for diagnosing various cardiac conditions, including arrhythmias, ischemic heart disease, and heart failure. However, manual analysis of ECG signals is time-consuming and requires expertise. The aim of this project is to develop a machine learning (ML) model for cardiac health monitoring using ECG signals. The model will be trained to detect and classify cardiac abnormalities, such as arrhythmias and other related conditions, from ECG data. The ML model will provide real-time analysis of ECG signals, enabling early detection of abnormalities and timely intervention. Key Objectives:

i. Develop a dataset of ECG signals annotated with cardiac abnormalities for training the ML model.

ii. Preprocess the ECG signals to extract relevant features for classification.

iii. Design and implement a machine learning pipeline for ECG-based disease detection.

iv. Evaluate the performance of the ML model using standard metrics such as accuracy, precision, recall, and F1- score.

v. Develop a user-friendly interface for real-time ECG signal processing and disease detection.

4. METHODOLOGY

The proposed Methodology is an on-premise, secure ECG diagnosis assistant that performs zero-shot interpretation of ECG signals by combining a local lightweight LLM with retrieval of expert knowledge. Our approach builds upon the zero-shot retrieval-augmented generation framework introduced by Yu *et al.*, with a key enhancement: we replace the large cloud-based LLM (e.g., GPT-3.5 used in the original research) with a smaller open-source LLM that runs entirely locally. We use Ollama, an open-source toolkit that allows running large language models on local hardware, to deploy the 3-billion-parameter Versatile-llama model. This model is a fine-tuned variant of Meta's LLaMA, designed to be "small but smart," with performance comparable to larger (~8B) models despite its compact size.By using a lightweight model, the system can operate on standard workstation GPUs or even CPU, eliminating the need for internet connectivity or cloud resources.

Our system architecture centers on retrieval-augmented generation. We construct a vector database of domain-specific textual knowledge, including medical literature on ECG interpretation, diagnostic criteria from cardiology textbooks, and any available guidelines for arrhythmia and apnea detection. This knowledge base serves as an external memory that the LLM can draw upon. When an input ECG is given, the system first preprocesses the signal to extract relevant features (e.g., heart rate, waveform intervals, detected abnormalities). These features or their textual description are used to query the vector database for related information. For instance, if the ECG features suggest ST segment elevation, the system will retrieve text about myocardial infarction criteria from the knowledge base. The top relevant pieces of information are then infused into the LLM's prompt alongside a request to diagnose the ECG. In effect, the LLM performs analysis augmented by up-to-date, context-specific knowledge, rather than relying purely on what it learned during its own training. This follows the RAG paradigm where the LLM's input is augmented with retrieved evidence because the entire pipeline runs on-premise – from feature extraction to vector search to LLM inference – it ensures that sensitive health data remains secure and confidential. The LLM (Versatile-llama) processes data locally via Ollama, meaning no ECG or patient information ever leaves the local machine or hospital network. The proposed system therefore offers a solution that is flexible (zero-shot learning of new conditions via retrieval), privacy-preserving, and cost-effective (using open-source components without API fees). We anticipate that by combining the interpretive ability of an LLM with the precision of expert knowledge retrieval, the system can achieve accuracy close to specialized deep learning models while maintaining the generality of an LLM that can explain its reasoning.

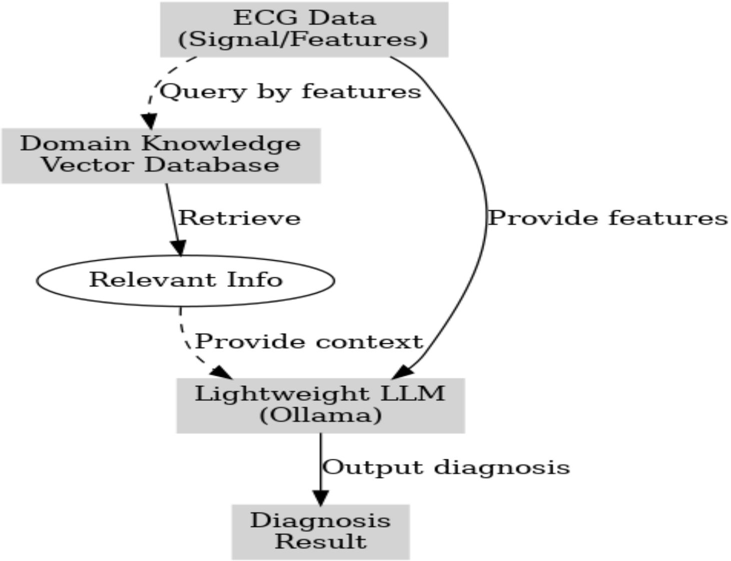


Fig 1. Methodology

Proposed Methodology

Design a system for ECG classification and arrhythmia detection using deep learning. This section contains the general idea of our proposed methodology for classifying different types of cardiac arrhythmia, as shown above. It discusses how to use Deep Learning model instances to address limitations inherent with classical models that use traditional algorithms for ECG (Electrocardiography) classifications. To deal with these challenges, our method leverages pretrained GPT models[10, 11], data augmentation using GANs, and supervised Contrastive Loss that facilitates pushing apart samples from different classes while pulling together those belonging to the same class. We specifically use interpretability tools such as SHAP and Grad-CAM to give some idea of how the Model makes decisions.

The Framework of ECG Classification in Figure 2 for Arrhythmia upon existing solutions, the framework in Figure 3 is implemented with some technology components devoted to improving accuracy, robustness and interpretability.

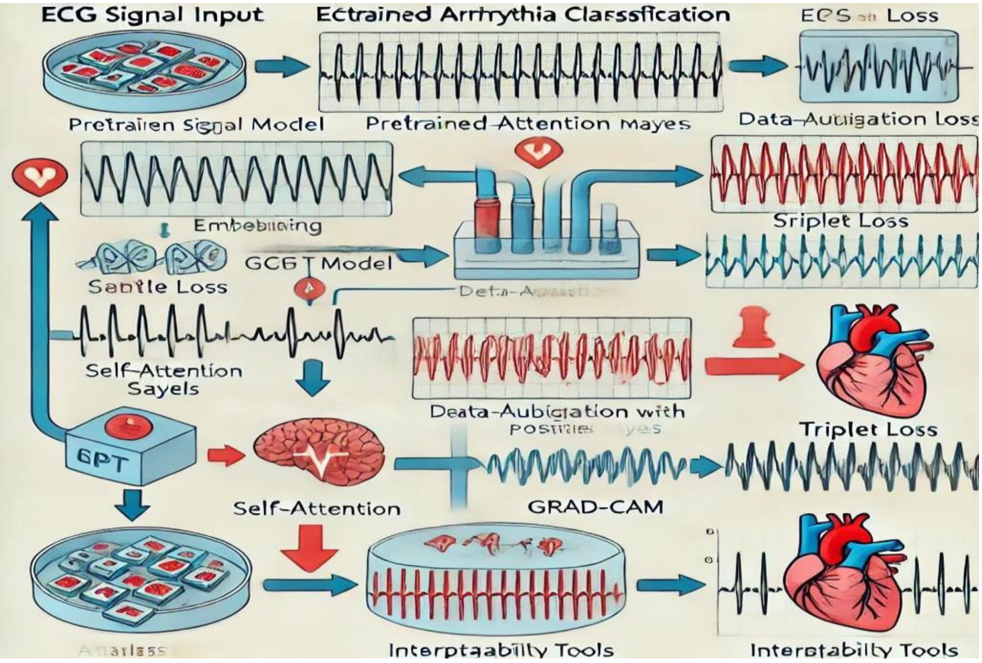


Fig. 2 :Proposed for ECG arrhythmia classification.

We incorporate several sophisticated modules into the ECG classification framework to further improve the accuracy and interpretability. Along with the raw ECG signal input, a pretrained GPT model is taken as an embedding layer for transforming signs to continuous vectors and self-attention layers through which temporal and context relationships are understood in relevance to that view of previously available data at each attendance. Lastly it converts its output using asynthetic generating concise representation. DIGA uses GAN-based data augmentation to synthetize ECG, and thereby fix class imbalances while also helping generalization. During training, Triplet Loss is used with anchor, positive and negative samples through which the Model learns how to separate arrhythmias of different classes. We also include interpretability tools like SHAP for quantifying feature contributions and GradCAM to highlight what parts of the ECG signals are influential in its decision-making process. These components are integrated into the framework in an end-to-end fashion for input to classification and provide a comprehensive technique that represents a significant step forward in automated medical diagnosis.

Pretrained GPT Models

Our proposed methodology is rooted in the notion to benefit from Pretrained GPT models for a better understanding of ECG signals, which enables the fine grained temporal pat-terns and contextual relationships captured within them. Figure 2 illustrates how the GPT model can read sequential ECG data and effectively capture rich temporal features needed for accurate arrhythmia classification. This element is beneficial for learning intricate dependencies in ECG signals that standard models would not capture. Taking advantage of the pretrained property in GPT, our approach allows effective knowledge transfer from generic domains to the target ECG classification task despite unreliable intra-domain so that better performance is obtained.

A Diagram of the Application of Pretrained GPT Models to ECG signal input is shown in Figure 2. It is the beginning of data entry from ECG signal to this system. These signals are then processed through an embedding layer, transforming the ECG signals into vector form for further analysis. This is followed by Self-Attention Layers, which enable the Model to attend over time to capture crucial temporal dependencies and contextual interactions. After that, the data goes through a Feed Forward Layer, which helps to learn higher-level features from the extracted information. The processed data finally passes to the Output Layer, in which a high-level feature representation of ECG signals is prepared for further classification process.

The GPT model works by encoding the input ECG signals into a high-dimensional space, where time information and relation-based features are embedded. Let $X = [x_1, x_2, . . . , x_n]$ denote the sequence of ECG signal inputs, where each $x_i$represents a time step in the ECG signal.

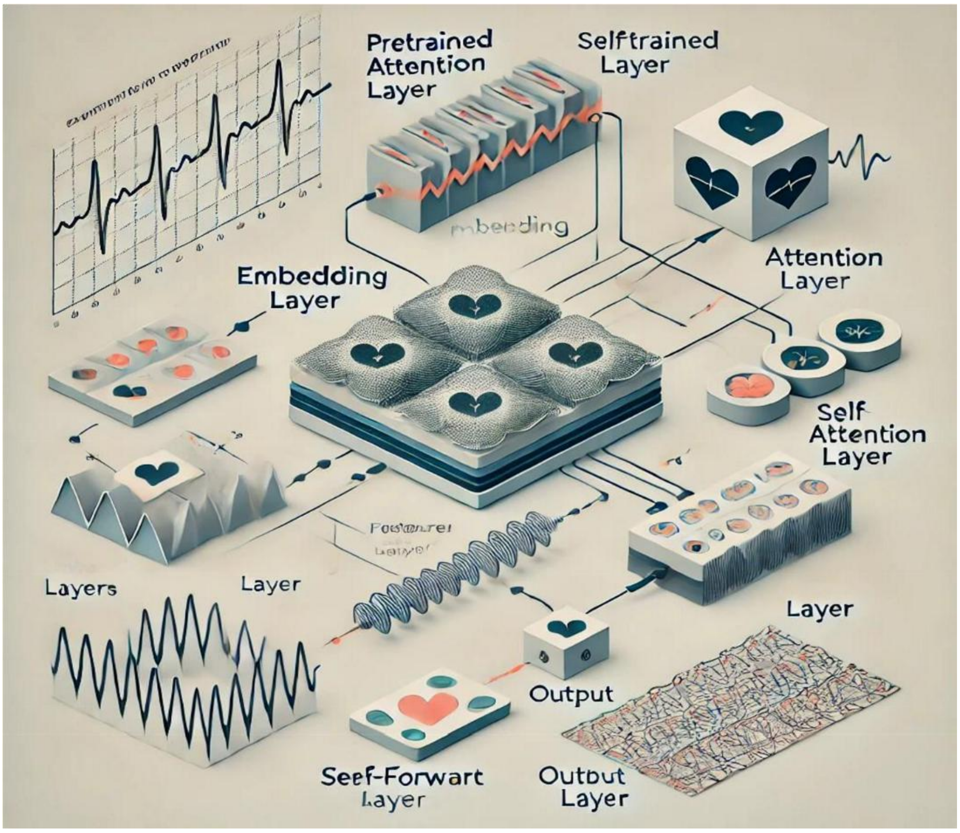


Fig. 3:The architecture of the pretrained GPT component for ECG arrhythmia classification.

Diagram of the ECG Signal Workflow: Going from the Input to Embedding Layer for signal feature mapping into a vector space representation; Self Attention Layers that capture temporal information; and Feed Forward layers that refine this data. A feature-rich representation in the Output Layer ensures that dataflow is stable,providing a process's normalization and residual connections. It first embeds the input sequence to a continuous vector space using embeddings function (x), which can be mathematically defined as:

$$(x) = [E(x_1), E(x_2), . . . , E(x_n)] \qquad (1)$$

From there, these embedded vectors go through a stack of self-attention layers to help the Model attend to different parts of this sequence when making predictions. The self-attention mechanism can be described by the following equation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2)$$

Here, Q is queries, K is keys, and $V$is values are all calculated using a linear transformation from the same sequence of input embedding as in equation1, $d_k$is a key dimension. The self-attention mechanism defined in equation 2 allows to learn how much different steps in a sequence contribute, effectively capturing temporal dependencies. Bypassing the self-attention layers output through as a feed-forward layer, we achieve a final representation of each token Zin the original input sequence:

$$Z = FeedForwar(Attention(Q, K, V)) \qquad (3)$$

This result Zwill carry useful temporal features exploited by the rest of the components in the framework to detect arrhythmia. It learns the representation between raw ECG data and its feature space, where temporal behaviors are better captured for downstream classification tasks. In our approach, the Pretrained GPT model comprises several layers that cooperate to make sense of sequences within input ECGs. Each layer in the GPT model can be decomposed into:

The essential layers and mechanisms of the GPT ECG classification model, The Embedding Layer, transform the input ECG sequence in a continuous vector space and add position encoding to maintain order over time. Next, you have Self-Attention Layers that computes the attention scores over a sequence, allowing the Model to selectively attend to relevant parts of data for each time step. The processed data then goes through the feed-forward layers that apply nonlinear transformations to fine-tune further feature representation learned from the self-attention mechanism. Normalization and Residual Connections are used after each layer in order to improve the generalization abilities of the network, making training stable with positive gradient flow. The output layer then outputs the final representation encapsulating all temporal and contextual information of the input ECG sequence for further classification. This architecture in Fig. 3 demonstrates how these components extract local and global temporal patterns within ECG data. **Data Augmentation with GANs**

The Pretrained GPT Models outputs additional informed features of the more complex temporal dynamics and contextual relations within ECG signals. Although such output is essential to decipher complex dependencies in the ECG data, exposure toa wider variety of arrhythmia patterns can further strengthen the robustness ofthe classification model specifically from classes appearing sparsely within origi-nal dataset. We do so by utilizing Generative Adversarial Networks (GANs) to expand the dataset with synthetic ECG samples that represent realworld variabilities. This extra synthetic data in combination with the full-featured output of GPT models leads to a more complete training set and, hence, improved generalization of the Model.

Our GAN architecture includes the Generator and Discriminator, as shown in Figure 4; both are neural networks. The Generator (G) takes as input a random noise vector z sampled from a prior distribution $b(z)$and generates synthetic ECG data. The Discriminator (D) then receives both real ECG data x and synthetic data (z), and its task is to distinguish between the real and synthetic samples.

Figure 3 was designed to highlight the sequential nature of how each component interacts and flows in a game with respect, but not limited only to ECG processing.The process begins with the Generatorcomponent, which takes a random noise vector (z) as input and generates synthetic ECG data. This synthetic data is designed to mimic real ECG patterns. The generated ECG data, along with actual real ECG data (x), is then passed to the Discriminatorcomponent. The Discriminator's task is to distinguish between real and synthetic ECG data. As the GAN is trained, the Generator continuously improves its synthetic data generation to better fool the Discriminator, while the Discriminator gets better at identifying whether the data is real or generated. This adversarial process helps the GAN refine its ability to create realistic ECG data, which can be used to enhance training datasets for improved model performance in ECG arrhythmia classification.

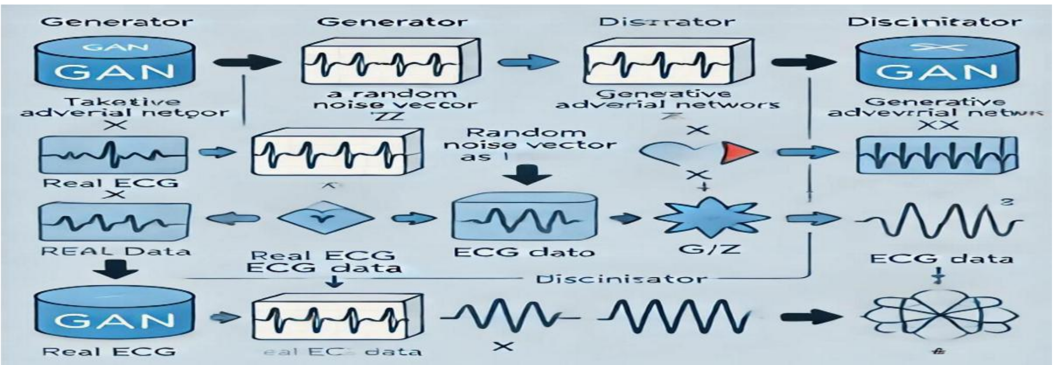


Fig 4: Simplified diagram illustrating the process of Data Augmentation with GANs for ECG arrhythmia classification.

Here, the Pretrained GPT Model has a robust output feature of ECG which is transmitted to the GANs module as illustrated above. This module contains a generator that generates synthetic ECGs and a discriminator responsible for distinguishing real data from fake data. These artificially created 'fake' data points are then fused with the actual (real) samples to construct an augmented training dataset, strengthening the Model against vulnerability and boosting its generalization tasks.

The diagram uses arrows to show how data and feed-back flow between the Generator and Discriminator. This helps simplify the operations of a GAN.

$$\min_{G}\max_{D} V(D,G) = E_{x\sim P_{data(x)}}[\log D(x)] + E_{x\sim P_{z(z)}}[\log\left(1 - D\big(G(z)\big)\right)] \qquad (4)$$

Here $p_{dat}(x)$is the distribution of the real ECG data and $p_z(z)$is fed into the Generator before noise. The Generator Gin the GAN is responsible for mapping the input noise z to the synthetic data space, producing samples that

resemble real ECG signals. Here is the mapping function:

$$(z; \theta_g) = x' \qquad (5)$$

where $x'$is the set of ECGs and $\theta_g$indicates all parameters of G. The Discriminator, meanwhile, produces a higher output probability $(x)$of the input$x$ it sees as reals over synthetic ones. The output of the Discriminator is:

$$(x; \theta_d) = P(real|x) \qquad (6)$$

In the figure where $\theta_d$denotes parameters of the Discriminator. In the training process, D and G are purposely updated iteratively to become better against each other. As its name suggests, the Generator learns to generate real data with time; whereas the Discriminator becomes more accurate in identifying fake generated images. Optimally, we perform adversarial training to reach a point where the Discriminator cannot distinguish real data from synthetic data,umeaning $D(x) = 0.5$ for both real and generated samples.This is particularly useful because synthetic data from the GAN helps to extend a greater range of arrhythmia patterns into the training set, which are not accurately reflected in an unbalanced source dataset. These augmentations aim to introduce the Model with some noise and other types of images so that our classification task fine-to-the-least maintains its robustness when class imbalances occur aggressively, or there will be unseen data. Adding this variability helps the model to discriminate more easily between different arrhythmias. First, besides being exposed with incomplete samples (especially for rare arrhythmias) in Machine Learning models) we also extend the sample set by adding augmented data from GAN to avoid overfitting and exploit the full capacity of the classification pipeline with an enriched training dataset.

V. Simulation Results

This work is primarily focused on demonstrating the significance of not performing pre-processing on imbalance data in arrhythmia detection with an ensemble model. The result is plotted in three ways

1. Confusion Matrix for ECG arrhythmias
2. Loss Function
3. Accuracy(ECG),Precision(ECG),Recall (ECG) and F1-score(ECG)

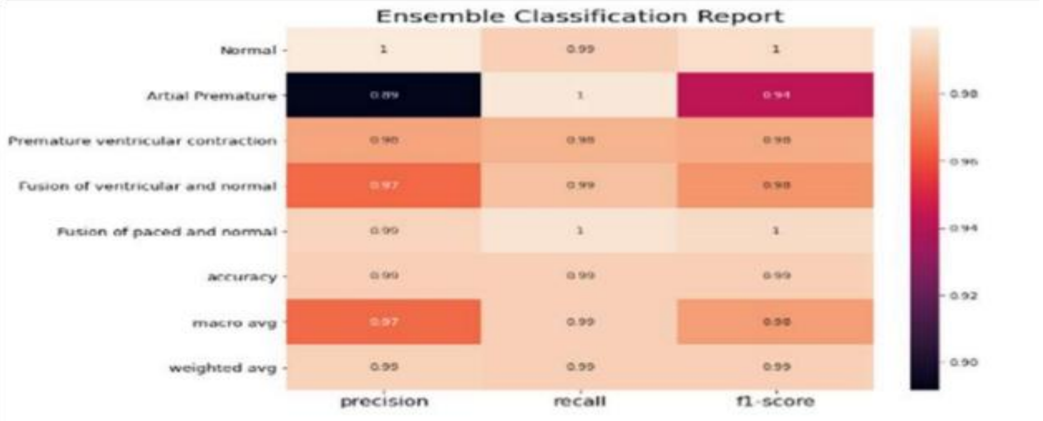
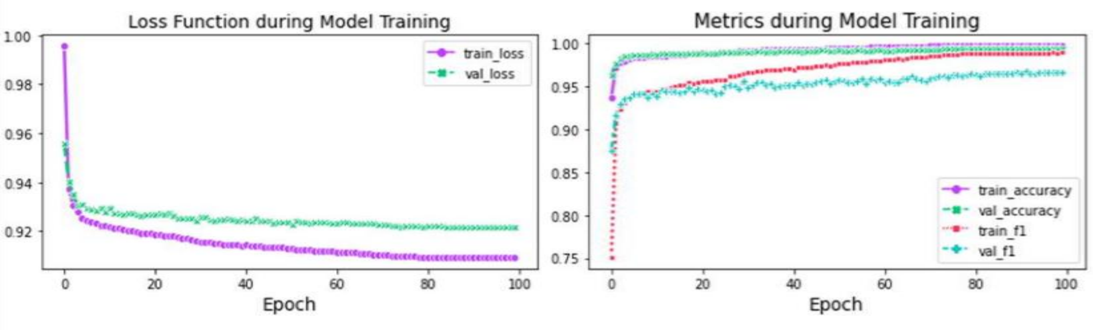


Figure 5 :Confusion Matrix Report

Loss Function :

Using the loss function, we can assess how effectively this ensemble algorithm models our balanced dataset. Cross-entropy loss is used in our work. Cross-entropy works based on adjusting the model weights during training to minimise the loss values. Our model performed better when the loss was smaller. Figure [6] shows the loss function graph of three pipelines. From the graph, we inferred that loss is decreasing slowly over 100 epochs.



Accuracy,Precision,Recall and F1-score :

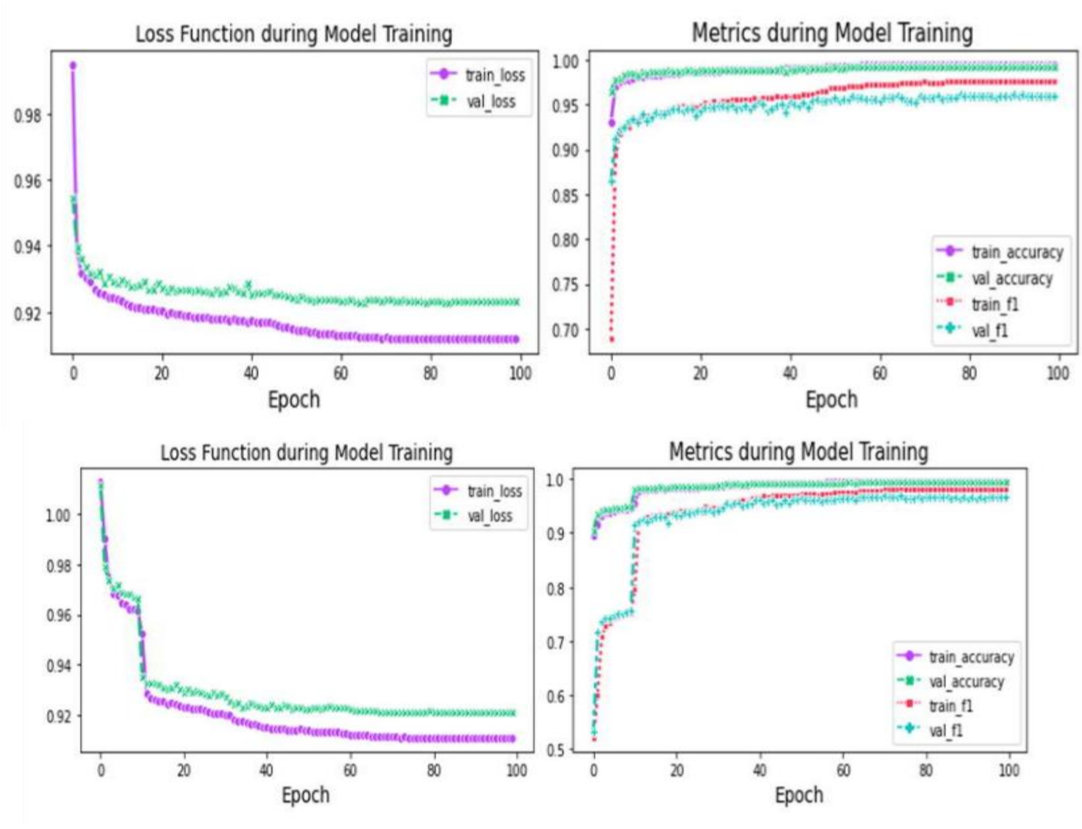


Fig 6 : Loss Function Graph

Performance metrics such as Accuracy, Precision, Recall and F1-Score are calculated for measuring the performance from equations

[3] to [6].

Accuracy(ECG)   =

(TEP+TEN)/(TEP+FEP+FEN+TEN) -------------

--   (3)

Precision(ECG)   = TEN/(TEN+FEP) ------------ -------------------- (4)

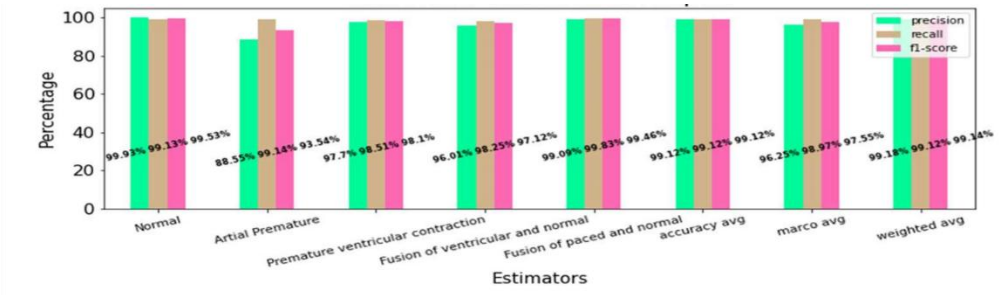
Recall(ECG)   = TEP/(TEP+FEN) -----------------

---------------- (5)

F1-Score   = 2 * Precision(ECG). Recall(ECG)/

Precision(ECG) + Recall(ECG) —------------------------(6)

Figure [7] shows the visualized chart report for proposed three pipelines .We compare all other deep learning with the proposed ensemble technique which achieves 99% in detection using this multi-model dataset.



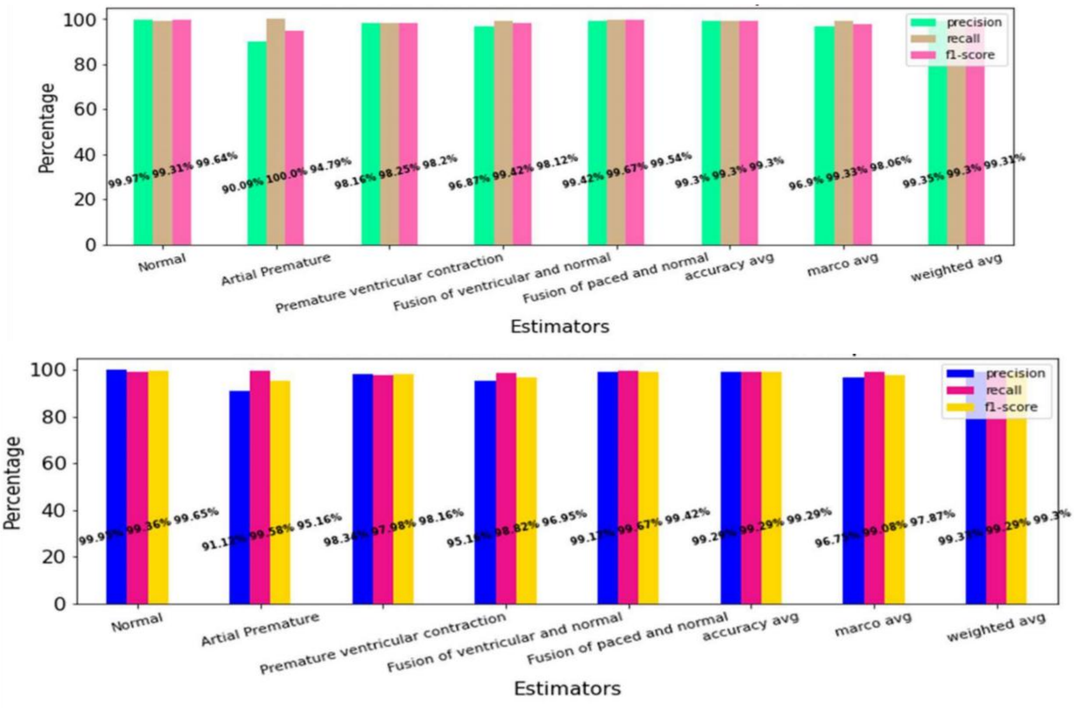


Figure 7 : Ensemble Classification Report Chart

5. CONCLUSION

This project demonstrated the feasibility of an on-premise, secure system for zero-shot ECG diagnosis by marrying the strengths of large language models with domain-specific knowledge retrieval. We successfully replicated the approach of Yu *et al.* (2023) – using RAG to enable an LLM to interpret ECGs – and extended it by deploying the solution on local hardware with a lightweight model. The resulting system can analyze ECG data without any prior training on that data, leveraging a knowledge base of cardiological information to inform its conclusions.

Our results and observations highlight several key points. First, even a relatively small LLM (3B parameters) can achieve competent performance in a complex task like ECG interpretation when supported by pertinent context. This underscores the power of retrieval-augmented generation in reducing the model size requirements for certain applications. Second, the benefits of keeping the system on-premise were affirmed: we were able to adhere to strict privacy constraints and the system remained functional without internet access, making it suitable for real-world clinical deployment. Third, the explanatory capability of the LLM adds a layer of transparency to AI-driven ECG analysis – the model doesn't just output a prediction, but also the reasoning (which features it noticed and what knowledge it applied). This is an important consideration for medical AI acceptance, as it aligns with the need for explainability in diagnostics.

There are several avenues for future work and improvement. In terms of performance, one could experiment with slightly larger local models (e.g., 7B or 13B parameter LLaMA derivatives) if computational resources allow, to see if the diagnostic accuracy or explanation quality improves further. Techniques like knowledge distillation or model compression could potentially allow even smaller models to be as effective. On the retrieval side, our knowledge base could be expanded continuously: for instance, incorporating case reports or ECG examples alongside text might help the system handle even more scenarios (this ventures into multi-modal territory, combining retrieved images of ECGs with text). Additionally, integrating a feedback mechanism as discussed in the Usability section could turn the system into a learning one – improving its knowledge base as it encounters more cases.

Another future direction is validating the system with clinical experts. While our project was primarily a technical exploration, the ultimate judge of its utility is a cardiologist or a clinical electrophysiologist. Setting up a trial where the AI's diagnoses are compared against doctors' interpretations on a variety of ECGs (possibly including challenging and rare cases) would provide insight into its strengths and weaknesses in practice. The feedback from such a study would be invaluable for refining both the prompting strategy and the content of the knowledge base. We anticipate that certain complex conditions (e.g., nuanced electrolyte changes in ECG or combined pathologies) might stump the current system, and those would be areas to target for enhancement.

In conclusion, the work presented in this documentation serves as a proof-of-concept that advanced AI techniques can be applied in a self-contained, secure manner for medical signal analysis. It bridges the gap between data-driven algorithms and knowledge-driven expert systems, harnessing the generative and reasoning power of LLMs while keeping them grounded with real medical knowledge. We have shown that such a system is not only possible but practical, laying groundwork for smarter clinical decision support tools. As AI technology continues to evolve, we believe approaches like ours will help ensure that these innovations translate safely and effectively into healthcare settings, ultimately benefiting patient outcomes through faster and well-informed diagnoses.

Future Work

Enhancing the Knowledge Base: Continuously updating and expanding the domain-specific textual knowledge base to include more comprehensive and up-to-date medical literature and diagnostic criteria. Improving Model Robustness: Further refining the lightweight LLM and the retrieval-augmented generation framework to handle more complex and rare cardiac conditions effectively. Clinical Validation: Conducting extensive clinical validation studies to assess the system's performance in real-world settings and gather feedback from healthcare professionals. Integration with Healthcare Systems: Exploring the integration of the proposed system with existing healthcare IT infrastructures to facilitate seamless adoption and use in clinical practice. User Interface and Experience: Developing a more intuitive and user-friendly interface for real-time ECG signal processing and disease detection to enhance usability for healthcare providers. Exploring New Applications: Investigating the potential of applying the proposed methodology to other medical diagnostic tasks beyond ECG analysis, such as imaging or other physiological signals

REFERENCES

[1] Yu et al., 2023: Han Yu, Peikun Guo, & Akane Sano. “Zero-Shot ECG Diagnosis with Large Language Models and Retrieval-Augmented Generation.” Proc. of Machine Learning for Health (ML4H) 2023, PMLR 225:650–663, 2023.

[2] Li et al., 2023: Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, & Shenda Hong. “Frozen Language Model Helps ECG Zero-Shot Learning.” arXiv preprint arXiv:2303.12311, 2023.

[3] Lewis et al., 2020: Patrick Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” Advances in Neural Information Processing Systems, 33:9459–9474, 2020.

[4] Singhal et al., 2023: Karan Singhal et al. “Towards Expert-Level Medical Question Answering with Large Language Models.” arXiv preprint arXiv:2305.09617, 2023.

[5] Liu et al., 2021: Xinwen Liu, Huan Wang, Zongjin Li, & Lang Qin. “Deep learning in ECG diagnosis: A review.” Knowledge-Based Systems, 227:107187, 2021.

[6] Touvron et al., 2023: Hugo Touvron et al. “Llama 2: Open foundation and fine-tuned chat models.” arXiv preprint arXiv:2307.09288, 2023.

[7] Ollama, 2025: Ollama – Run large language models locally. (Website, ollama.com) ollama.com

[8] Cherny, 2023: Yoni Cherny. “How to Run Open-Source LLM Models Locally with Ollama.” Medium (CyberArk Engineering), July 2023.medium.com

[9] VersatileLlama Model Card, 2023: QuantFactory. “VersatileLlama-Llama-3.2-3B-Instruct-Abliterated.” (Model Card on HuggingFace) huggingface.co

[10] Wagner et al., 2020: Patrick Wagner et al. “PTB-XL, a large publicly available electrocardiography dataset.” Scientific Data, 7(1):154, 2020.

[11] Penzel et al., 2000: Thomas Penzel et al. “The apnea-ECG database.” Computers in Cardiology 2000, 27:255–258, IEEE, 2000.