

Train model report

Classifying Arrhythmia in Fighter Pilots

APRIL 4, 2025

Group 3

Picture 147, Picture

Contents

[Introduction 2](#_Toc701849547)

[1.Model Performance 3](#_Toc1990554762)

[2.Model Selection 6](#_Toc1198237034)

[3. Model Implementation 10](#_Toc1853847254)

[References 13](#_Toc1553969603)

# Introduction

Electrocardiogram (ECG) signals are vital for diagnosing cardiac abnormalities such as arrhythmias. Automating this classification process can improve early detection and reduce the burden on healthcare professionals. In this project, we investigate machine learning and deep learning methods to classify ECG signals into two categories: normal and abnormal heartbeats.

Due to the highly imbalanced nature of the dataset, where normal beats dominate, model evaluation requires careful attention. This report presents a systematic approach to selecting and implementing an effective classification model, starting with a basic logistic regression and culminating in a more advanced hybrid model that combines Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks.

The report begins with an analysis of model performance, where the results of logistic regression are compared against the baseline accuracy, and the challenges caused by class imbalance are discussed. This is followed by model selection, which provides a comparison between different models suitable for ECG classification and explains the motivation behind choosing the hybrid CNN + LSTM architecture. Finally, the model implementation section outlines the preprocessing steps, architectural design, training process, and parameter choices used to build the final model, accompanied by code snippets and visuals for clarity.

Together, these sections highlight the rationale behind our approach and provide a comprehensive view of how deep learning techniques can be effectively applied to ECG classification tasks.

# 1.Model Performance

In binary classification tasks, the baseline accuracy refers to the performance achieved by always predicting the most frequent class in the dataset. It serves as a simple benchmark to evaluate whether a trained model provides any added value.

In our case, the dataset consists of two classes:

* **N** (Normal)
* **A** (Abnormal)

Out of all instances:

* **98.5%** belonged to class **'N'**
* **1.5%** belonged to class **'A'**

If we were to always predict 'N', we would achieve a baseline accuracy of: 98,5%

However, this strategy would completely fail to detect abnormal heartbeats (class 'A'), which are the most critical cases to identify in medical diagnostics.

We used Logistic Regression as a baseline model. The data was split as follows:

* 70% for training
* 15% for validation
* 15% for testing (with stratification to preserve class distribution)

The model was trained with class\_weight="balanced" to compensate for the class imbalance.

After training, the model achieved an accuracy score of 0.9971 on the test set.

While this result appears excellent, it must be interpreted with caution. The dataset is heavily biased toward normal heartbeats, and the test set contained very few abnormal samples. As a result, the model achieved a high score which mostly reflects the correct classification of normal heartbeats.

The initial performance metrics on the test set showed as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| Normal (N) | 1.00 | 1.00 | 1.00 | 336 |
| Abnormal (A) | 1.00 | 0.80 | 0.89 | 5 |
| Accuracy |  |  | 0.9971 | 341 |

These metrics show that while the model performed well on normal beats, it missed 1 out of 5 abnormal cases, which is crucial in real-world ECG monitoring scenarios. This underlines the limitations of accuracy alone and the importance of recall and F1-score for minority classes.

To assess generalizability, we applied 5-fold cross-validation. This method helps identify overfitting or fluctuations in performance across different data splits, especially for the rare class 'A'.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Fold | Accuracy | Precision (A) | Recall (A) | F1-Score (A) |
| 1 | 0.9934 | 0.83 | 0.71 | 0.77 |
| 2 | 0.9978 | 1.00 | 0.86 | 0.92 |
| 3 | 0.9978 | 1.00 | 0.83 | 0.91 |
| 4 | 0.9956 | 0.86 | 0.86 | 0.86 |
| 5 | 0.9934 | 1.00 | 0.57 | 0.73 |
| Average | 0.9956 | 0.94 | 0.77 | 0.84 |

The model consistently achieved high accuracy (~99.5%) and excellent precision for abnormal heartbeats across all folds. However, recall varied more significantly, ranging from 0.57 to 0.86. This variation reveals an important insight: while the model is highly reliable when it predicts an abnormality (high precision), it sometimes fails to detect abnormal heartbeats at all (lower recall).

This is especially critical in medical applications, where missing an arrhythmia can have serious consequences. These findings highlight the limitations of relying on accuracy alone and support the need for exploring more advanced models, such as CNNs and LSTMs, which are better equipped to capture subtle and complex patterns in ECG data — particularly those present in the minority class.

Areas for Improvement

* **Dataset Imbalance**: The extreme class imbalance limits the model’s ability to learn and detect abnormal heartbeats effectively. Collecting more samples of abnormal heartbeats (class 'A') would improve the model’s ability to generalize and make reliable predictions across both classes.
* **Data Augmentation**: Techniques like SMOTE or synthetic ECG generation could help balance the dataset and reduce overfitting to the majority class.

These findings highlight the limitations of relying on accuracy alone and emphasize the importance of models that can better capture rare but clinically important patterns. Based on this insight, we next explore and compare several machine learning and deep learning architectures to identify a more effective solution.

# 2.Model Selection

For this project, we explored multiple models to classify ECG signals effectively. The following models were trained and evaluated:

* **Logistic Regression**
* **Convolutional Neural Network (CNN)**
* **Long Short-Term Memory Network (LSTM)**
* **Hybrid CNN + LSTM model**

Model Comparison:

1. Logistic Regression

We began with Logistic Regression as a baseline model due to its simplicity and interpretability. It assumes a linear relationship between input features and the output class.

Pros:

* Fast to train and easy to implement
* Requires fewer computational resources
* Provides a clear baseline for comparison

Cons:

* Assumes linearity, which is often unrealistic for time-series biomedical data
* Limited ability to capture complex patterns in ECG signals
* Struggles with imbalanced datasets without careful tuning

1. Convolutional Neural Network (CNN)

Next, we implemented a Convolutional Neural Network, a popular architecture for processing sequential and spatial data like ECG signals. CNNs are effective at automatically extracting local features through filters and pooling layers.

Pros:

* Captures local patterns such as sharp peaks or dips in ECG signals
* Learns features directly from raw or minimally processed input
* Performs well on structured data like 1D signal sequences

Cons:

* Requires more data and training time than classical models
* May miss long-term dependencies across the signal without additional layers

1. Long Short-Term Memory (LSTM)

We also tested a Long Short-Term Memory (LSTM) network, a type of recurrent neural network (RNN) designed to capture temporal dependencies in sequential data.

Pros:

* Effective at modeling time-based dependencies in ECG data
* Handles variable-length input sequences
* Complements CNNs by analyzing sequential patterns over time

Cons:

* Computationally heavier and slower to train
* Requires more careful tuning and regularization

1. Hybrid Model (CNN + LSTM)

For the final model, we selected a hybrid CNN + LSTM architecture, combining the strengths of both approaches:

* The **CNN** extracts local features from the ECG signal
* The **LSTM** captures temporal relationships between those features over time

This hybrid approach is well-suited for ECG classification, as it can both detect important waveform shapes and understand their sequence within the heartbeat signal.

To objectively compare the models, all of them were assessed using a 5-fold cross-validation. The following table summarizes the average performance across folds, focusing on metrics most relevant to medical anomaly detection: precision, recall and F1-score for the minority class. This method provides a consistent basis for comparison and avoids biases that might result from a single train-test split.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision (A) | Recall (A) | F1 Score (A) |
| Logistic Regression | 0.9956 | 0.94 | 0.77 | 0.84 |
| CNN | 0.9938 | 0.97 | 0.62 | 0.73 |
| LTSM | 0.9850 | 0.20 | 0.03 | 0.06 |
| CNN + LTSM | 0.9850 | 0.00 | 0.00 | 0.00 |

The table above summarizes the performance of each model, with an emphasis on metrics relevant to the minority class — abnormal heartbeats (A). Logistic Regression, although a simple model, achieved strong and consistent performance with a recall of 0.77 and F1 score of 0.84, indicating a solid ability to detect abnormalities across folds.

The CNN model reached the highest precision (0.97) for abnormal cases, showing it made very few false positive predictions. However, its recall was lower (0.62), suggesting it often missed actual abnormal cases — a critical issue in medical diagnostics.

The LSTM model performed poorly on the minority class, with an average F1 score of just 0.06, indicating its limited effectiveness under the current setup. Most notably, the CNN + LSTM hybrid model failed to detect any abnormal cases across all folds, resulting in zero precision, recall, and F1 score for class A. This reflects the challenges these deep learning models faced in learning from a highly imbalanced dataset without additional class-balancing techniques.

These findings reinforce the importance of model simplicity in the presence of class imbalance. Logistic Regression, in this context, provided the most reliable performance, while more complex models may require further tuning, class weighting, or data augmentation to be viable alternatives.

Model Selection

Despite its failure to detect abnormal heartbeats in our current evaluation, the CNN + LSTM model was selected due to its architectural strengths and support in scientific literature. The hybrid architecture combines CNN’s strength in extracting spatial features from ECG waveforms with LSTM’s ability to learn temporal dependencies, a synergy that aligns well with the sequential nature of ECG data.

Research by Rahul & Sharma (2025) and Denysyuk et al. (2023) demonstrates the potential of such hybrid models, particularly when supported by larger datasets and appropriate class-balancing techniques. Additionally, in high-performance aviation contexts such as fighter pilot monitoring, the ability to capture non-linear, high-frequency variations is essential. Soh et al. (2024) note that physiological monitoring in these environments benefits from more advanced AI models capable of adapting to extreme conditions.

Therefore, while the CNN + LSTM model underperformed in our current configuration, it remains a strategically sound choice for future development and deployment, particularly with further optimization, data augmentation, and class imbalance mitigation.

Discussion and future work

While this project demonstrated promising results, particularly with Logistic Regression, the selected CNN + LSTM hybrid model faced notable challenges in detecting abnormal heartbeats. These limitations offer valuable insight into the constraints of deep learning when working with imbalanced medical datasets. The table below summarizes the key issues encountered during model development and evaluation, along with recommended strategies for addressing them in future iterations of this project.

|  |  |  |
| --- | --- | --- |
| Challenge | Impact | Suggested improvement |
| Class imbalance | Model biased toward majority class | Apply class weighting, oversampling (e.g., SMOTE), or generate synthetic ECGs |
| CNN+LSTM failed to detect class A | Zero recall and F1-score for minority class | Tune hyperparameters, apply focal loss, and extend training time |
| Small dataset size | Limited learning, especially for deep models | Collect more real-world abnormal ECG data or use data augmentation techniques |

Addressing these limitations would likely improve the deep learning models’ ability to detect arrhythmias, especially in complex or real-time scenarios such as those involving high-performance fighter pilots.

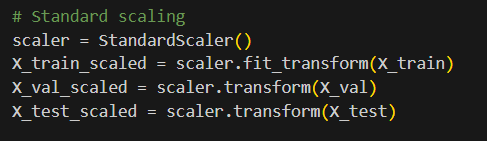
Other studies, such as Daydulo et al. (2023), have also demonstrated the viability of deep learning approaches for arrhythmia detection using time-frequency features, reinforcing the potential of CNN and LSTM-based architectures when supported by appropriate signal representation and tuning.

# 3. Model Implementation

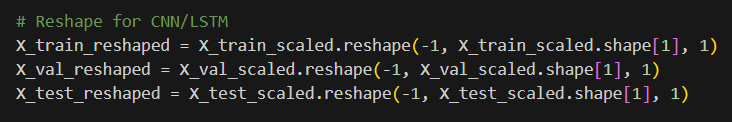
In this chapter, we describe the steps taken to implement the chosen hybrid CNN + LSTM model for ECG classification. The implementation was carried out in Python using TensorFlow/Keras, with attention to code clarity, data preprocessing, and appropriate parameter tuning.

Data Preprocessing

The ECG dataset was first standardized to ensure all features had zero mean and unit variance. This step is crucial for neural networks, as unscaled input data can negatively impact model convergence.



After scaling, the data was reshaped to fit the CNN + LSTM input format. Each sample was reshaped to a 2D array (samples, time\_steps, channels).

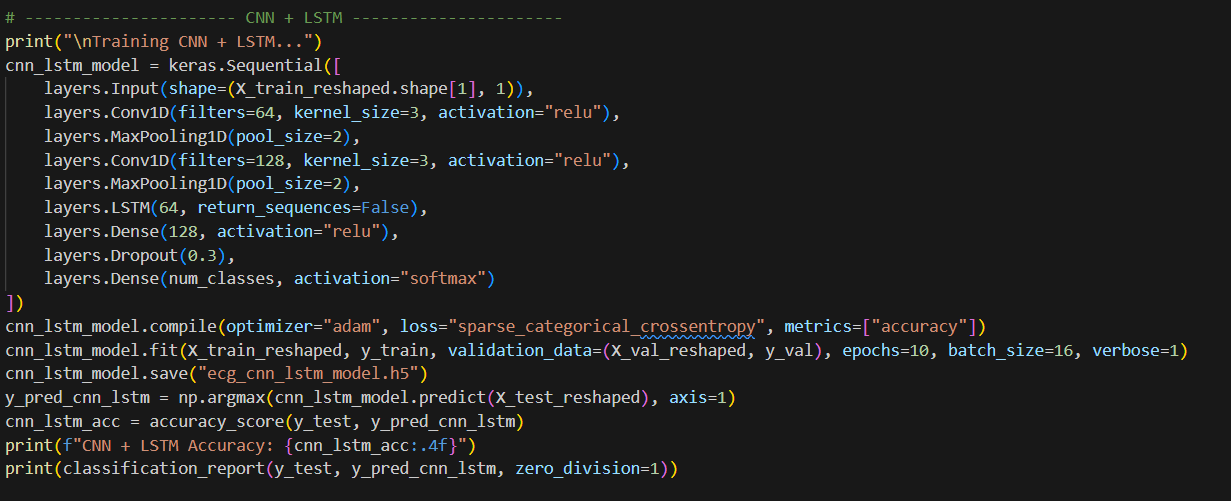


Neural networks like CNN and LSTM expect input in the format (samples, time\_steps, channels). This reshaping step prepares the data accordingly, where each heartbeat sample becomes a 2D time series with one feature channel.

Model Architecture

The hybrid model consists of:

1. 1D Convolutional layers to extract local ECG patterns
2. MaxPooling layers to reduce dimensionality and focus on important features
3. LSTM layers to capture temporal dependencies in the ECG sequences
4. Dense output layer with a sigmoid activation for binary classification



Parameter Choices

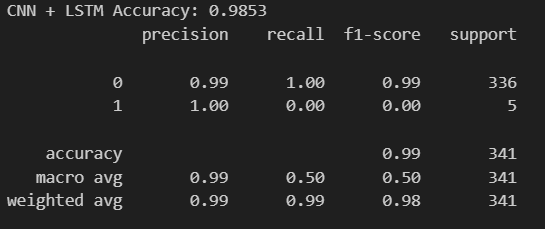
Several parameters were tuned during the experimentation process. These included:

|  |  |  |  |
| --- | --- | --- | --- |
| **Component** | **Parameter** | **Value** | **Reason** |
| Conv1D | Filters | 64 | Balanced feature richness and speed |
| Conv1D | Kernel Size | 3 | Small enough to capture sharp waveform changes |
| LSTM | Units | 64 | Adequate capacity to model temporal dependencies |
| Optimizer | Adam | lr = 0.001 | Adaptive learning, good for non-stationary inputs |
| Epochs | 10 | Prevent overfitting due to small dataset |  |
| Batch Size | 16 | Tradeoff between speed and performance |  |
| Loss Function | Binary Crossentropy | Standard for binary classification |  |



Training and Validation

The model was trained using the training set (70%) and validated on a separate 15% validation set. Accuracy, precision, and recall were tracked to monitor performance, especially for the minority class.



# References

*Rahul, J., & Sharma, L. D. (2025). Advancements in AI for cardiac arrhythmia detection: A comprehensive overview. Computer Science Review, 56, 100719.* [*https://doi.org/10.1016/j.cosrev.2024.100719*](https://doi.org/10.1016/j.cosrev.2024.100719)

*Hanna Vitaliyivna Denysyuk, Rui João Pinto, Pedro Miguel Silva, Rui Pedro Duarte, Francisco Alexandre Marinho, Pimenta, L., António Jorge Gouveia, Norberto Jorge Gonçalves, Coelho, P., Eftim Zdravevski, Petre Lameski, Reis, V., Garcia, N. M., & Ivan Miguel Pires. (2023). Algorithms for automated diagnosis of cardiovascular diseases based on ECG data: A comprehensive systematic review. 9(2), e13601–e13601.* [*https://doi.org/10.1016/j.heliyon.2023.e13601*](https://doi.org/10.1016/j.heliyon.2023.e13601)

*Yared Daniel Daydulo, Bheema Lingaiah Thamineni, & Ahmed Ali Dawud. (2023). Cardiac arrhythmia detection using deep learning approach and time frequency representation of ECG signals. BMC Medical Informatics and Decision Making, 23(1).* [*https://doi.org/10.1186/s12911-023-02326-w*](https://doi.org/10.1186/s12911-023-02326-w)

*Soh, M.-S., Jang, J.-H., Park, J.-S., & Shin, J.-H. (2024). Effects of high-gravity acceleration forces and anti-gravity maneuver on the cardiac function of fighter pilots. Scientific Reports, 14(1), 8749.* [*https://doi.org/10.1038/s41598-024-59274-2*](https://doi.org/10.1038/s41598-024-59274-2)