Final report: Heartbeat Classification

by *Christian Meister*
Julia Schmidt
Tzu-Jung Huang

20. November 2025

Introduction:

Motivation

Heart diseases are currently a leading cause of death globally [1]. While Electrocardiograms (ECGs) are the standard tool for diagnosing heart conditions [1, 2], manual analysis is time-consuming and prone to human error. Therefore, developing automated systems to classify heart signals could be crucial for medical professionals to improve patient outcomes by facilitating the fast initiation of medical treatments.

Project Goals

This project focused on the automation of two critical tasks based on ECG data:

1. **Arrhythmia Classification:** Categorizing 5 types of heartbeats using the MIT-BIH dataset.
2. **Myocardial Infarction (MI, Heart Attack) Detection:** Distinguishing between healthy and abnormal heartbeats using the PTB dataset.

Our primary objective was to build a model that outperforms the 2018 benchmark study by Kachuee et al. [2].

Our Approach

We faced significant challenges regarding the class imbalance present in both datasets, which required us to move beyond standard modeling techniques. To address these issues, we implemented a robust pipeline that included advanced **data augmentation** to handle dataset limitations and **Transfer Learning** to apply knowledge from models trained on a larger dataset (MIT-BIH) to models trained on smaller ones (PTB), while retaining time efficiency for the modeling part.

Key Achievements

The results of our approach were successful. We surpassed the performance of the benchmark paper:

- **MIT-BIH Dataset:** Our optimized **CNN model** achieved an accuracy of **98.51%**.
- **PTB Dataset:** Our transfer learning model achieved an accuracy of **98.42%**.
- **Reliability:** We further validated our models using **SHAP analysis**, confirming that the predictions are based on medically relevant signal features.

Difficulties encountered during the project

- **What was the main scientific obstacle encountered during this project?**

Class imbalance and sampling strategy:

The MIT dataset suffers from severe class imbalance, with normal beats constituting over 82% of the data. We initially tested Random Oversampling, but rejected this method as it merely duplicates data, leading to an overfitting problem. Consequently, we adopted SMOTE as our standard sampling approach, as the performance of baseline models was the best using this method while preventing overfitting. SMOTE generates synthetic samples rather than duplicates, providing a more robust solution for minority classes (like class 3) and ensuring consistent model performance.

Handling R distance extreme values for class 0 of the MIT dataset:

In Rendering 1, we identified the R distance for every sample in the dataset. For class 0, we observed that data belonging to extreme values of the R distance tended to exhibit signal patterns that, from our perspective, did not appear normal. As medical experts annotated the dataset, and removing samples belonging to those extreme values did not lead to an improvement of the model performance (neither for baseline models nor for DL models), we had to rely on the annotations made. However, it could be helpful for future projects to verify the signals belonging to the R distance extreme values by medical experts to exclude possible present misclassifications.

Interpretation of the DL results:

Balancing model performance with clinical interpretability was also challenging. While DL models achieved superior performance, their 'black box' nature conflicts with the clinical need for transparent, explainable predictions that medical experts could verify and trust in critical decision-making scenarios. With the SHAP analysis,

it was possible to gain insights into the decision-making of the used models. However, it remains the most challenging part of the project for us as the patterns can be very complex and difficult to identify.

- **For each of the following points, if you encountered difficulties, detail how they slowed you down in setting up your project.**
- **Forecast: tasks that took longer than expected, etc.**

Hyperparameter tuning and model optimization required iterative refinement that extended beyond our original schedule. Ultimately, many processes were running simultaneously, which made it difficult to keep track of everything.

- **Datasets: acquisition, volumetry, processing, aggregation, etc.**

The present class imbalances were challenging. Initially, we sought an oversampling method that would achieve the best performance. However, the results had to be critically analyzed, as the best performance was achieved with the RandomOversampler. After further discussion, this method was evaluated as inappropriate, and SMOTE was selected as the most suitable approach for our use case. Testing different baseline models with different oversampling methods was very time-consuming, even though only RandomizedSearch was performed for this test.

- **Technical/theoretical skills: timing of skill acquisition, skill not offered in training, etc.**

Accurate evaluation and decision-making regarding which DL model is most suitable based on loss and accuracy curves throughout the training process. It was challenging to understand when loss and accuracy curves indicated sufficient convergence and if we should include early stopping in our code. We needed to discuss best practices in evaluating the training procedure and model performance in specific cases. E.g., that convergence of the loss/accuracy curves is the most important point, and that it is best practice always to use the model of the last epoch. This could be clarified in the training.

- **Relevance: of the approach, model, data, etc.**

We had several different ideas for the modeling phase. Especially when comparing the use of raw ECG signals versus handcrafted features (such as P-wave duration and QT interval), it is essential to discuss which approach is the most clinically relevant. We tested various methods, but due to time limitations, we were unable to test every idea; therefore,

we ultimately focused on the raw ECG signals in the end, as this is the data directly available in clinical contexts.

- **IT: storage power, computational power, etc.**

Extensive hyperparameter tuning for the baseline models (especially using GridSearch) required significant processing power and took a long time. Training DL models was also time-consuming, which forced us to abandon complex (and time-consuming, given the modeling procedure's requirement) CNN-LSTM hybrid architectures and to concentrate only on testing a few different model types and architectures to stay on schedule. To resolve this bottleneck, we migrated our training workflow to Google Colab using GPUs, which significantly accelerated computation and enabled our models to converge within the project timeline.

- **Other**

Organizing and managing various tasks, while maintaining an overview of ongoing processes and upcoming tasks.

Report

- **Detail what your main contribution was to achieving the project's goals.**

Christian:
- Literature search
- First dataset analysis
- Baseline modeling
- Optimization of baseline models
- Interpretation of results
- DL Model double-check
- Merging Code Structures

Julia:
- Literature search
- First dataset analysis
- DL modeling
- Optimization of DL models
- Interpretation of results

Tzu-Jung (Kiki):
- First dataset analysis
- Managing different tasks
- Model check and verification
- Interpretation of results
- Integrator

- **Have you changed the model since the last iteration? If yes, provide details.**

No, we did not change the model.

- **Present the results obtained and compare them to the benchmark**

As the best-performing DL model was inspired by the architecture presented in [2] and used the same datasets, the results are compared with those in [2] (see Table 1).
For the model trained on the MIT dataset, only the average accuracy was presented in the paper. The best DL model found outperforms the model in [2], achieving an average accuracy of 0.9851, which is ~5% higher than the result in [2].
For the transfer learning model retrained on the PTB dataset, the average accuracy, precision, and recall were given in [2]. The performance of the best DL model found in this project was higher for all the given metrics. An average accuracy of ~2.5% higher, an average precision of 2.3% higher, and an average recall of more than 3% higher could be

achieved. It is worth noting that in [2] all residual blocks of the CNN were frozen during the transfer learning procedure. Here, we adapted the transfer learning procedure and retrained the CNN with the last residual block unfrozen. With this approach, it was possible to achieve a higher performance than in [2].
The found DL models outperformed the model performances achieved in [2].

		Metrics on test data			
Model	**Dataset**	**Avg Accuracy**	**Avg Precision**	**Avg Recall**	**Avg F1 Score**
[2]	MIT	0.9340	-	-	-
CNN8	MIT	0.9851	0.9062	0.9424	0.9236
[2]	PTB	0.9590	0.9520	0.9510	-
CNN8 + transfer6	PTB	0.9842	0.9751	0.9864	0.9805

Table 1 - Result summary of performance achieved in [2] and performance of the best found DL models.

- **For each of the project's goals, detail how they were achieved or not.**

Find the best-performing baseline model and appropriate oversampling technique:

- RandomizedSearch: get first impression of model performance with different sampling techniques
- GridSearch: optimize the best model found in RandomizedSearch with the most appropriate sampling technique

Find DL model architecture with best performance:

- Test different architectures presented in scientific papers or in lectures

Achieve better performance than in [2] with the DL model:

- Test different DL model versions with added dropout and batch normalization layers
- Test different training setups: optimize batch size and learning rate
- Transfer learning: unfreeze the last residual block to improve performance

Interpretation:

- SHAP analysis to identify patterns in the ECG signal that are important for the decision-making of the model

- **If they have been reached, in which process(es) can your model fit? Detail.**

- Arrhythmia classification based on ECG signals
- MI detection based on ECG signals

Continuation of the project

- **What avenues for improvement do you suggest to increase the performance of your model?**

It would be important to increase the quality of the datasets regarding class balance.
For the MIT dataset, especially for classes 1 and 3, additional data should be added. In addition, for class 0, the present labels should be reviewed by medical experts, particularly for samples with extreme R distance values (as identified in Rendering 1). Some of the ECG signals contained signal parts that did not look normal from our perspective. It would be interesting to test whether changes like this further increase the model's performance and minimize misclassifications. This would be especially important for misclassifications belonging to true labels 1-4, which are predicted as class 0, to reduce the number of false negative predictions. These are particularly dangerous in the application of the method in clinical contexts, as this could lead to emergencies being missed.
For the DL model retrained on the PTB dataset, only a few misclassifications were present. The pre-trained model compensated for class imbalances present in the PTB dataset and the smaller number of present samples. Therefore, the priority would be to increase the dataset quality for the MIT dataset.

- **How has your project contributed to an increase in scientific knowledge?**

An automated ECG signal analysis pipeline could be developed for clinical contexts to classify ECG signals as normal or abnormal while providing initial identification of specific arrhythmias or MI using trained DL models like those presented in this project. Medical experts would then verify these preliminary assessments. Such a system would be valuable for obtaining rapid initial evaluations of patient status, potentially accelerating clinical decision-making and treatment initiation.

It is essential to emphasize that the model’s results should serve as a decision-making tool, rather than a definitive diagnosis. In-depth review and validation by qualified medical professionals is not only clinically valuable but also ethically essential to ensure patient safety and maintain an appropriate standard of medical care.

Bibliography

- **What bibliographical elements (research articles, blogs, books, etc.) did you rely on to carry out your project?**

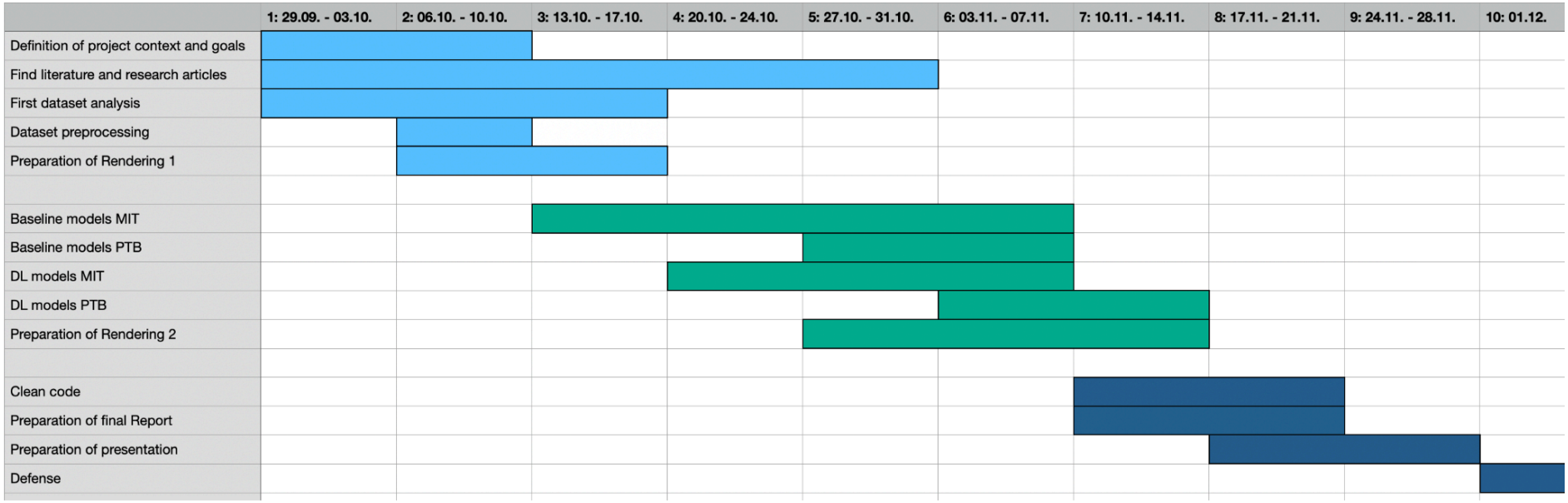
We mainly used research articles to get inspiration for DL model architectures or for baseline models that are popular for arrhythmia classification or MI detection.

Most important research articles:

- **ECG Heartbeat Classification: A Deep Transferable Representation**; M. Kachuee, S. Fazeli, M. Sarrafzadeh (2018); *CoRR;* doi: 10.48550/arXiv.1805.00794
- **Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review**; F. Murat, O. Yildirim, M, Talo, U. B. Baloglu, Y. Demir, U. R. Acharya (2020); *Computers in Biology and Medicine*; doi:10.1016/j.compbiomed.2020.103726
- **Deep learning for ECG Arrhythmia detection and classification: an overview of progress for period 2017–2023**; Y. Ansari, O. Mourad, K. Qaraqe, E. Serpedin (2023); doi: 10.3389/fphys.2023.1246746

Appendices

- **Gantt Diagram**



- **Description of code files.**

Dataset analysis and preprocessing:

- Class distribution
- R distance analysis
- Preprocessing: delete duplicates for the PTB dataset, training/validation/test dataset generation

Baseline modeling:

- Without oversampling
- With oversampling: Pipeline

DL modeling:

- DNN
- CNN
- LSTM
- Transfer

SHAP analysis:

- CNN
- CNN transfer

Citations

[1] Deep learning for ECG Arrhythmia detection and classification: an overview of progress for period 2017–2023; Y. Ansari, O. Mourad, K. Qaraqe, E. Serpedin (2023); doi: 10.3389/fphys.2023.1246746

[2] ECG Heartbeat Classification: A Deep Transferable Representation; M. Kachuee, S. Fazeli, M. Sarrafzadeh (2018); *CoRR;* doi: 10.48550/arXiv.1805.00794