Detecting Heartbeat Normality and Abnormality Using Machine and Deep Learning Techniques

Team members: Ping Yuan[1], Dr. Sepideh Tabrik[2], Isabell Gurstein[3]

[1] ping.yuan.paris@gmail.com

[2] sepideh.tabrik@gmail.com

[3] i.gurstein@gmail.com

Mentor: Francesco Madrisotti

GitHub repository: https://github.com/Ping-YUAN/heartbeat-analysis-ai

Abstract:

In this report, we use advanced machine learning and deep learning techniques to detect whether a heartbeat is normal or abnormal. It aims to improve the accuracy of Heartbeat detection and classification within general supervision by using both traditional machine learning techniques, as well as where more advanced Deep Learning models are powerful. We provide a comprehensive overview of the methodologies employed, including data exploration, data preprocessing, feature engineering, model selection, and evaluation metrics. Conclusion Overall our results show promising efficacy of these techniques to improve diagnostic accuracy and possibly help in early diagnosis of cardiac anomalies. Supported by expert mentorship, our group work has resulted in detailed discussions and encouraging outcomes that could serve as groundwork for further research on medical diagnosis.

Introduction:

Electrocardiogram (ECG) signals are important diagnostic tools for monitoring a patient's heart health. These signals reflect the electric activity of the heart and can provide vital information about both normal and pathological heart functioning. Normal ECG signals show a particular pattern that represents every phase of a cardiac cycle, but deviations from this pattern may mean something is not right or there exists a possible abnormality in the heart. Abnormal ECG signals present as waveform anomalies, absence or extra beats, or irregular rhythms among others. Such abnormalities could indicate heart failure, arrhythmias, coronary artery disease, and other conditions.

We aimed to develop an intelligent model to classify these ECGs into two groups-normal and abnormal. By doing so, we might be able to develop a tool that might be useful in earlier detection of cardiovascular diseases hence improving patients' outcomes.

To begin with we explored and collated our dataset before any necessary pre-processing was done on it. Afterward, we evaluated our model's performance following its training using various machine learning algorithms and deep learning techniques. ECG signal manual analysis can take a long time and

DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

demands a high level of skill. This is where deep learning (DL) and machine learning (ML) come in handy.

ML and DL models are the best suited for assessing ECG signals because they can be trained to identify patterns within large datasets. For numerous benefits, develop a model that differentiates normal from abnormal ECG signals. With the evaluation of ECG signals by ML and DL models rather than humans, much data is treated quickly. Comparatively, machine learning (ML) and deep learning (DL) algorithms outperform manual analysis when equipped with proper training data. Continuous monitoring of ECG signals using ML and DL tools helps detect cardiovascular problems before they turn into serious health issues. Also, the development of these models on different platforms like smartphones and wearables will make cardiac monitoring more accessible.

Conclusively, there is much promise for using ML and DL in ECG signal processing for heart health monitoring as well as disease diagnosis purposes. It is also a milestone towards identification and intervention for various medical cases.

The ongoing Heartbeat Analysis project is an international collaboration that transcends business environments and national borders. Our team consists of Ping Yuan from France, Dr. Sepideh Tabrik, and Isabell Gurstein from Germany, each with their own set of viewpoints and skills.

Objective:

The main objectives of the project are to design, train, and optimize deep learning models for analyzing and classifying heartbeat sounds. Additionally, the project aims to provide team members with the opportunity to learn and apply deep learning techniques, equipping them with the skills to implement these technologies in their own professional fields. Furthermore, the project seeks to facilitate international collaboration, leveraging diverse expertise to enhance the overall outcomes.

Ping Yuan works at Datategy, a startup that builds daily tools for data scientists. With about 8 years of development experience, he has rich expertise in implementing features in code. He is interested in data science topics. His goal is to enhance his knowledge of mathematics to improve his understanding and application of data science concepts in daily life. He hopes to have the opportunity to switch his career path to data science after these classes. He has built a front-end to visualize heartbeats to assist in medical analysis. However, there are no AI or Machine Learning tools to help with predictions. He hopes that with this project, we can build a Docker image containing the model we built, along with a server and UI, so that medical professionals can load data through the UI and receive diagnostic assistance.

Dr. Sepideh Tabrik, throughout her academic journey, has acquired a broad knowledge in two distinct fields of study. She focused on developing methods for breast cancer detection using mammography during her master's degree studies. In her Ph.D. study, she concentrated on the neuroscience of neuroimaging, specifically how the brain perceives objects across different sensory modalities and exhibits plasticity in its response to these stimuli. She used functional magnetic resonance imaging (fMRI) data to explore how the brain processes and classifies stimuli from two sensory system modalities. Her fMRI dataset allowed her to acquire expertise in fMRI data acquisition, preprocessing,

DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

and analysis and an understanding of cognitive and sensory processing within the brain. She is currently on maternity leave but expects to gather new knowledge about data science that will enable her transition into industrial employment in Germany. Since she is from a medical background she would like to be part of the current project. This position would allow her to apply what she knows and build her skills, especially deep learning abilities.

Isabell Gurstein works as an SAP BI/Data Analyst at an SAP consulting company in Germany, specializing in the retail sector. Her goal is to transition into a data scientist role and eventually a machine learning engineer position within her current company. Although her current work primarily revolves around the retail industry, this project aligns with her original background in healthcare. She chose this project because it offers a unique opportunity to apply deep learning models to the analysis of sounds, allowing her to gain complementary skills that she can later implement in her company. With several years of experience in medical invoice auditing, two years as a research associate in geronto-psychiatric clinical trials, and an additional two years in IT project management at a health insurance company in Germany, Isabell has developed a strong expertise in the healthcare field and statistical analysis. She has also completed several courses in Machine Learning as part of her Master of Science degree, positioning her as a late beginner to early intermediate in machine learning proficiency. In her current role within the AI team, Isabell carries out machine learning algorithms and provides the chatbot with Business Intelligence (BI) data from the SAP context, enriching the bot's answers and enhancing its functionality. Despite her involvement in this advanced work, she has not yet consulted the machine learning expert on her team regarding this project. Through this initiative, Isabell aims to broaden her company's spectrum of clients and methodologies by integrating advanced data science techniques. She is excited about the potential to apply what she learns in this healthcare-focused project to innovate and expand the services offered by her company in the retail sector and beyond.

Rendering 1: exploration, data visualization, and data pre-processing report

These procedures are critical for converting raw data into useful insights and actionable information. This study will provide a detailed review of the techniques and methodologies used to explore datasets, illustrate data trends, and prepare data for further research.

Framework:

In this experiment, we utilize the freely available Heartbeat dataset from Kaggle. This dataset, curated by Shayan Fazeli, has a significant number of ECG signals classified into various classifications depending on their properties. The major classifications are 'Normal' and many types of 'Abnormal' heartbeats.

We simultaneously worked on two ECG Heartbeat Categorization Datasets obtained and cleaned beforehand for analysis. The first dataset originated from the MIT-BIH Arrhythmia Database (mitbih), while the second was based on the PTB Diagnostic ECG Database (ptbdb). Both the original and cleaned datasets are freely accessible, facilitating easy training and testing of our model on the prepared datasets and validation on the raw ones.

DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

MIT-BIH Arrhythmia Database includes 48 half-hour ECG recordings from 47 individuals, recorded between 1975-1979. 23 recordings were randomly selected from 4000, representing a mix of hospital patients and visitors. The other 25 recordings were chosen for their rare but important heart conditions. Each recording was digitized at a rate of 360 samples per second with 11-bit resolution within a 10 mV range. Over 110,000 annotations were made by cardiologists, providing a reference for each heartbeat. Two databases are available at https://www.kaggle.com/datasets/shayanfazeli/heartbeat, mitbih_train.csv, and mitbih_test.csv.

The mitbih_train.csv file is a crucial part of the ECG Heartbeat Categorization Dataset used for training models to recognize different types of heartbeats. Here are the detailed aspects of this file:

- Size and Structure: The file is quite large, with a size of 392 MB. It contains 87,554 rows, each representing a single heartbeat.
- Columns: There are 188 columns in the file. The first 187 columns represent the signal values at different time points, and the last column is the target label for the heartbeat class.
- Classes: The target column categorizes heartbeats into five classes:
    - 0: Normal heartbeats ("Normal")
    - 1: Supraventricular premature beats ("Supraventricular")
    - 2: Ventricular escape beats ("Ventricular")
    - 3: Fusion of ventricular and normal beats ("Fusion")
    - 4: Unclassified beats ("Unclassifiable")
- Class Imbalance: The dataset is known to be imbalanced, with a significant majority of the heartbeats labeled as normal (class "0"). This imbalance needs to be addressed during the model training process to prevent bias.

The mitbih_test.csv file is the testing counterpart to the mitbih_train.csv file in the ECG Heartbeat Categorization Dataset. Here are the details:

- Size and Structure: The file is approximately 98.1 MB in size. It contains 21,892 rows, with each row representing a single heartbeat.
- Columns: Similar to the training file, there are 188 columns. The first 187 columns represent the individual time points of the ECG signal, and the final column is the class label[2].
- Classes: The class labels are the same as in the training set, with five categories ranging from normal heartbeats to various types of arrhythmias.
- Usage: This file is used to evaluate the performance of the model trained on the mitbih_train.csv data. It helps in assessing how well the model generalizes to new, unseen data.

For PTB Diagnostic ECG Database a specialized recorder with 16 input channels was used, including 14 for ECGs, one for respiration, and one for line voltage. The recorder could handle an input voltage of ±16 mV, with a high resolution of 16 bits and a bandwidth of 0-1 kHz. It contains 549 records from 290 subjects, with ages ranging from 17 to 87. Signal Measurement: Each record has 15 signals, including the standard 12 ECG leads plus 3 Frank lead ECGs, all digitized at 1000 samples per second.

The ptbdb_normal.csv and ptbdb_abnormal.csv files are part of the PTB Diagnostic ECG Database, which includes ECG recordings from both healthy subjects and patients with various heart conditions. Here's a detailed explanation of these files:

DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

ptbdb_normal.csv: This file contains ECG recordings from healthy subjects. The data is structured similarly to the MITBIH files, with each row representing a single heartbeat and columns representing the signal values at different time points. The last column typically contains the label, which in this case would be 0 indicating a normal heartbeat.

ptbdb_abnormal.csv: This file includes ECG recordings from patients with various heart conditions. The structure is the same as the ptbdb_normal.csv file, with the last column containing the label 1 to indicate an abnormal heartbeat.

Given the two distinct types of datasets, we will describe each one separately to prevent any confusion.

Relevance:

Both datasets comprise 188 features, where features 0-186 constitute the data points for a single heartbeat, and the last feature represents a label. After data exploration, we have decided to consider all classes from 1 to 3 as abnormal heartbeats. However, the 'Unclassifiable' beat adds complexity without providing additional value for classification. Consequently, we have opted to exclude this class from the modeling process.

This dataset is a classification problem that uses supervised learning algorithms on the above-mentioned variables. The major aim of this classification is to differentiate between healthy and unhealthy heartbeats in order to develop models.

The major attributes of the dataset are as follows:

- ECG Data Points (Variables 0-186): These represent the amplitudes of ECG signals for one heartbeat. It is primarily characterized by the analysis and classification of heartbeats based on these parameters. Since all explanatory variables are sequential and represent specific segments of the heartbeat sequence for each case, they are uniformly important in predicting the target variable. As a result, no individual feature can be identified as particularly significant.
- Labels (Variable 187): This indicates how the heartbeats are classified thereby being necessary for supervised learning tasks (0: Normal, 1:Abnormal).

Data Limitations:

However, there are some limitations with this clean and well-structured dataset such as: for instance, this dataset lacks frequency components that can yield more information about ECG signals. Also lacking in patient metadata include age, medical history, etc.; all these details could be helpful in making personalized assessments. Furthermore, there is a data imbalance between normal and abnormal heartbeats which results in bias when implementing classification models.

1 - Pre-processing and feature engineering of MIT datasets:

Given that the initial data was already clean and all features were on a consistent scale, we carried out minimal feature engineering, specifically, converting the target variable type from float to numeric and

DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

adding labels, as the data was initially unlabeled. This allowed us to swiftly advance to the preprocessing stage. As the initial step in data visualization, we plotted the raw ECG signals for each distinct class—Normal, Unclassifiable, Ventricular, Supraventricular, and Fusion—for both the training dataset (Fig. 1) and the test dataset (Fig. 2).. This visualization aids in understanding the variations and patterns in heartbeat data, which are crucial for developing accurate machine learning models for heartbeat classification.

To further refine and finalize the previous step, when running histograms over a section of data points in the mitbih training dataset by each class, we realized that the shapes of heart beat classes do indeed differ from each other, as shown in the example of normal and supraventricular heart beats (Fig3a-b).

The distribution of different heartbeat categories in the training data and the test data is depicted in Figure 4 and Figure 5. This figure highlights the prevalence of each category, which is crucial for understanding the dataset's composition. This analysis reveals that the majority of the data consists of normal heartbeats, while the other categories are significantly less frequent. Understanding this distribution is essential for developing and training machine learning models, as it impacts the model's ability to accurately classify and detect abnormalities in heartbeats.
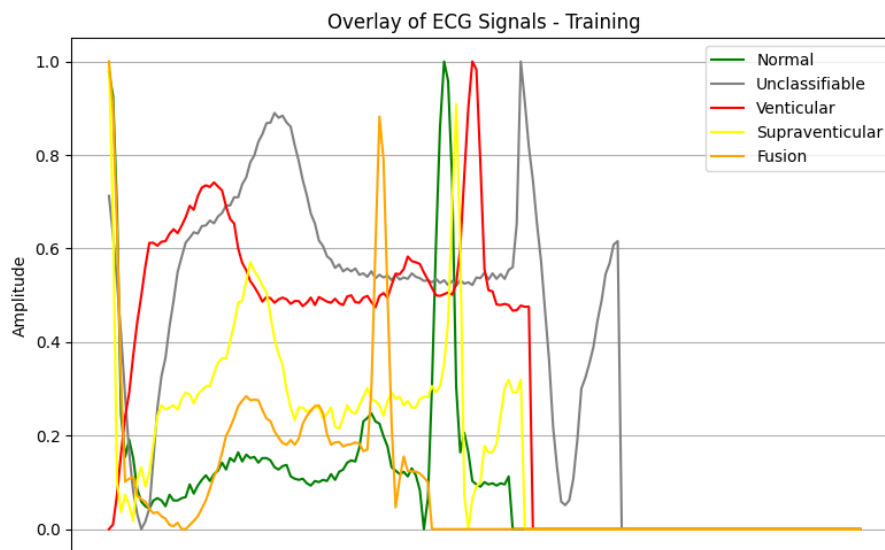

Fig1 - This graph illustrates the variations in ECG signals across different heartbeat categories during the training phase. The categories include normal (green), unclassifiable (grey), ventricular (red), supraventricular (yellow), and fusion (orange). Each line represents the unique electrical activity associated with each type of heartbeat, providing a visual comparison crucial for analyzing and understanding heart rhythms.
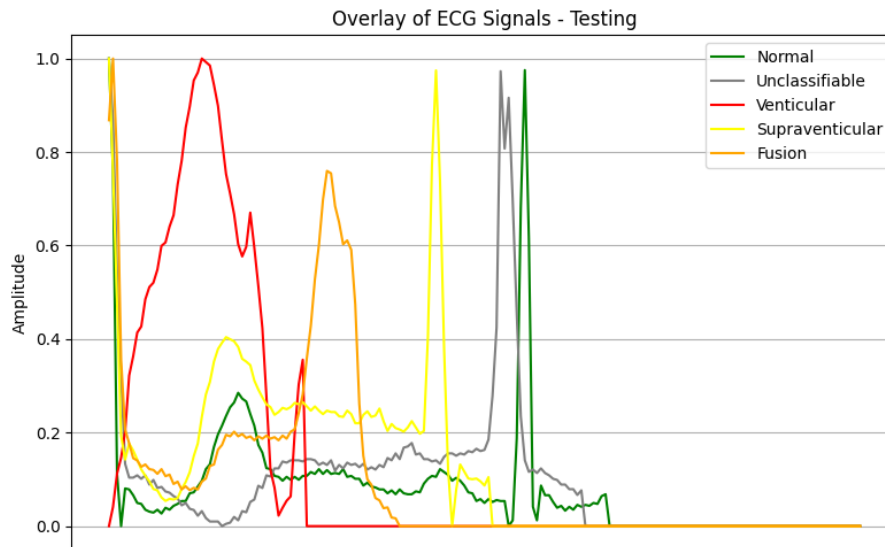
DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

Fig2 - This graph illustrates the variations in ECG signals across different heartbeat categories during the test phase. The categories include normal (green), unclassifiable (grey), ventricular (red), supraventricular (yellow), and fusion (orange). Each line represents the unique electrical activity associated with each type of heartbeat, providing a visual comparison crucial for analyzing and understanding heart rhythms.
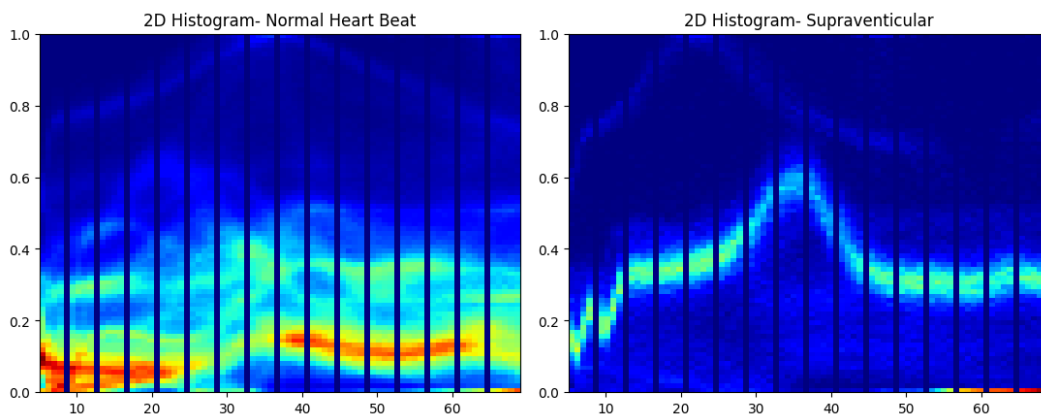

Fig3 - Two-dimention histograms over a section of data points [5-70] in the training dataset for (a) Normal class and (b) Supraventricular class.
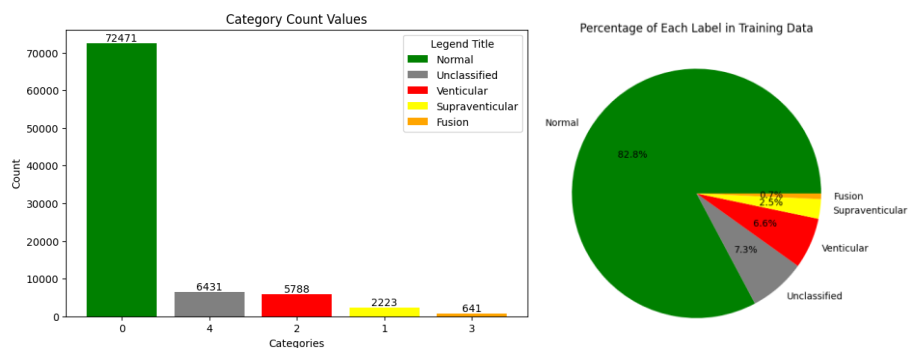

Fig4 - These bar and pie graphs display the count of values for different heartbeat categories in the training data. The categories include Normal (green) with 72,471 counts, Unclassified (grey) with 6,431 counts, Ventricular (red) with 5,788 counts, Supraventricular (yellow) with 2,223 counts, and Fusion (orange) with 641 counts.

DataScientest.com
Training organization approval 11755665975
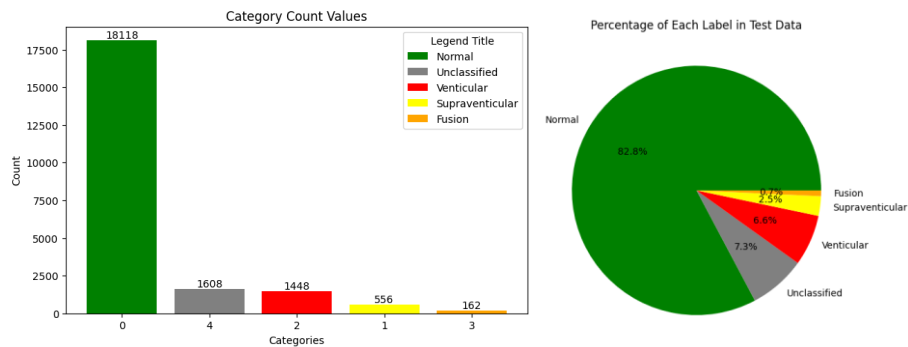09 80 80 79 49
2 place de Barcelona, 75016 Paris

Fig5 - These bar and pie graphs display the count of values for different heartbeat categories in the training data. The categories include Normal (green) with 18118 counts, Unclassified (grey) with 1608 counts, Ventricular (red) with 1448 counts, Supraventricular (yellow) with 556 counts, and Fusion (orange) with 162 counts.

PCA (Principal Component Analysis) is a powerful technique in the field of statistical analysis, often used to make sense out of complex data and present it more simply. This is done, for example using PCA (Principal Component Analysis), and in the field of interpreting heartbeat rate material helps visualization to distinguish between normal behavior and that of failing abnormalities.

PCA processes the original high-dimensional data to a lower-dimensional space enabling us to project our data into principal components revealing the highest significant structures. This simplification makes it easier to find clusters and outliers that represent different types of heartbeats (e.g., normal, supraventricular, ventricular beats) including fusions. In our case, we used PCA to a dataset including heartbeat rates, and the dimensions were reduced up to two principal components. This reduction enabled us to create a scatter plot that visually represents the distribution of different heartbeat types.
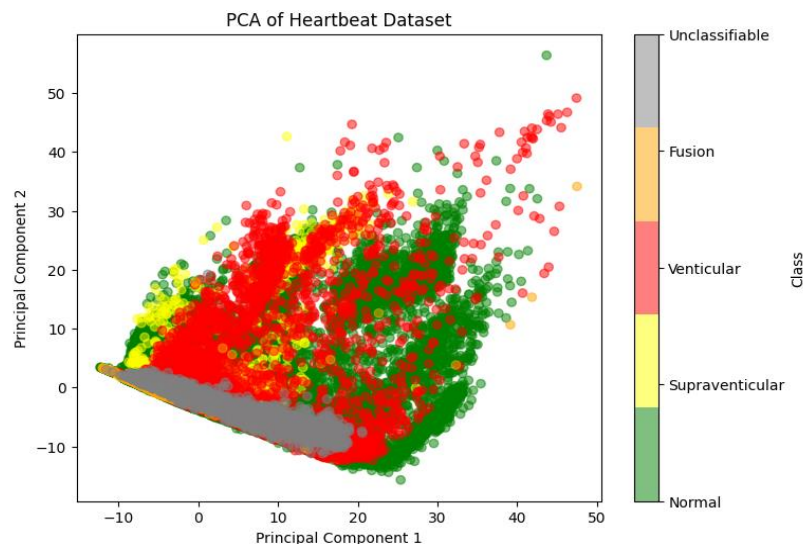

Fig6 - Scatter plot illustrating the distribution of heartbeat data projected onto the first two principal components after PCA dimension reduction. Each point represents an individual heartbeat, color-coded by class: Normal (green), Supraventricular (yellow), Ventricular (red) Fusion (orange), and Unclassifiable (grey). This visualization aids in distinguishing between normal and abnormal heartbeats.

DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

The plot in Figure 6 shows five distinct clusters for each heartbeat class. A higher density indicates more consistency within each class. For example, the Supraventricular heartbeats (yellow) form a dense cluster, suggesting consistent patterns in this heart activity. Additionally, we can see a significant overlap between the 'Normal' (green) and all other abnormal heartbeats, indicating that these classes share similar characteristics, which might make them harder to distinguish. This exploratory data analysis revealed that the classes of the target variable were not distinctly separated. Specifically, 'Fusion' (depicted in orange) represents a mixed category of 'Supraventricular' (yellow) and 'Ventricular' (red), while 'Unclassifiable' (gray) may potentially encompass elements from all other classes. Therefore, we should find a robust model to accurately detect normal and abnormal heartbeats. It is worth noting that points that are far from any cluster, such as some 'Ventricular' (red) heartbeats, can be considered outliers. These outliers might represent rare or unusual heartbeat patterns that could be clinically significant. As we have no information about the data frequency, we will keep all of them in the first step.

Reducing our data to two dimensions helps in visualizing the structure more clearly. To determine the optimal number of dimensions for representing our data, a scree plot is employed. A scree plot is a useful tool in Principal Component Analysis (PCA) for determining the number of principal components to retain. The plot displays the variance explained by each principal component, helping to identify the point where adding more components contributes little additional variance. In the scree plot (Fig 7), the x-axis represents the principal components, while the y-axis shows the percentage of variance explained by each component. The plot typically exhibits a steep decline initially, followed by a leveling off. The "elbow" point, where the slope changes most dramatically, indicates the optimal number of components to retain. From the scree plot, we observe a noticeable elbow after the second principal component. This suggests that the first two components capture the majority of the variance in the data, making them sufficient for further analysis. Retaining more components beyond this point would contribute minimally to explaining additional variance.
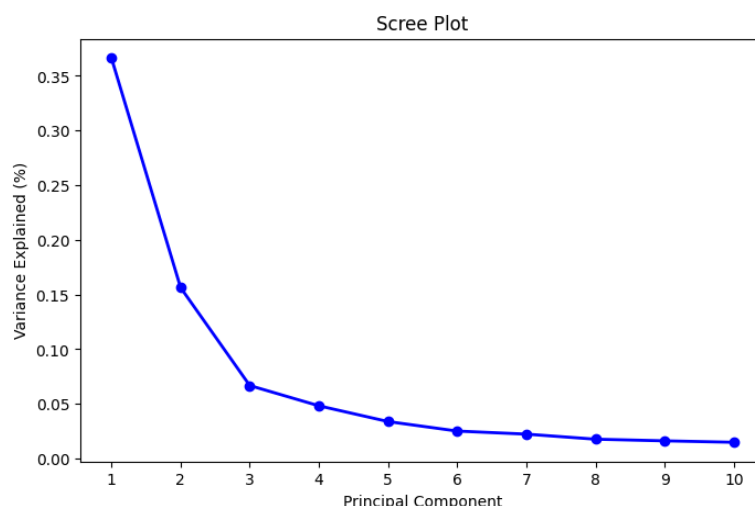

Fig6 - Scree plot illustrating the variance explained by each principal component. The elbow point after the second component suggests that retaining the first two principal components is optimal for capturing the significant variance in the dataset.

DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

Although the elbow in the scree plot suggests three dimensions, we chose to present our data in two dimensions. This decision enhances the clarity of our data representation in the report.

In addition, we have used a t-SNE technique to visualize the heartbeat classes. This map helps us see patterns in normal and abnormal heartbeats by grouping similar ones together. In other words, t-SNE is particularly useful for visualizing high-dimensional data in 2D or 3D, making it easier to see the natural groupings in the data. Each color on the map indicates a particular type of heartbeat, making it easier to determine which are distinct groups and which overlap, making model prediction more challenging.

In our context, clusters represent different classes of heartbeats (normal, ventricular, supraventricular, fusion, and unclassifiable). The distance between clusters indicates how distinct different types of heartbeats are from each other. Well-separated clusters suggest clear differences between the types. As it has been shown in Figure 7, there is separation and clustering between the abnormal classes but they are not dense and spread and somehow they overlap in some data points. As we can see there is also a huge data set from normal heartbeats that leads to overlap. The separation and clustering of these points can help identify specific patterns and anomalies in the heartbeats but the overlapping between them might lead to a challenging work on this project.
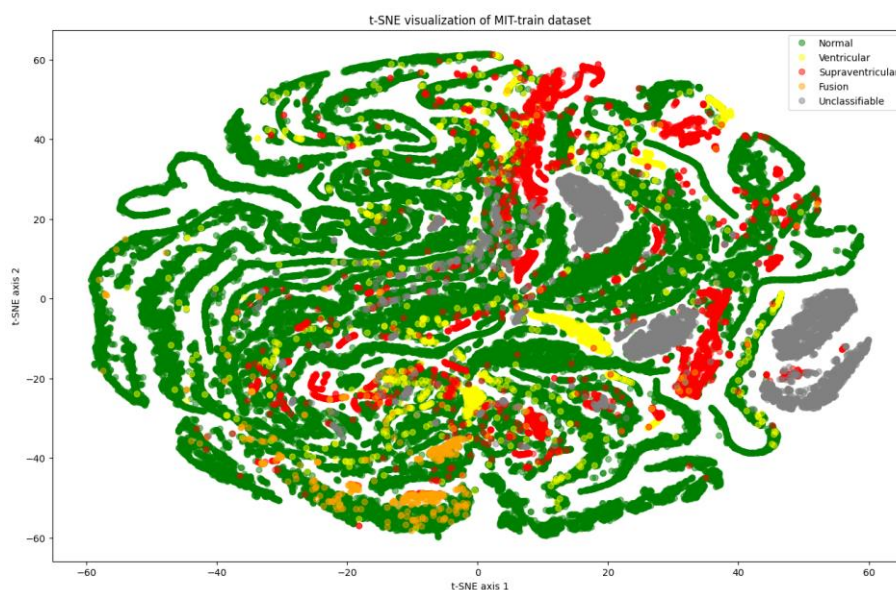

Fig 7 - t-SNE visualization of heartbeat classes from the MITtrain dataset. Green signifies normal heartbeats, red highlights Supraventricular abnormalities, yellow captures ventricular abnormalities, orange marks fusion beats, and grey indicates unclassifiable beats.

By using UMAP, we can gain valuable insights into the patterns of heartbeats, helping to identify and understand normal and abnormal conditions more effectively. As we can see in Figure 8, we have again 5 different classes, but they are not dense and the data points of each class are somehow spread, and clusters also overlap.

DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
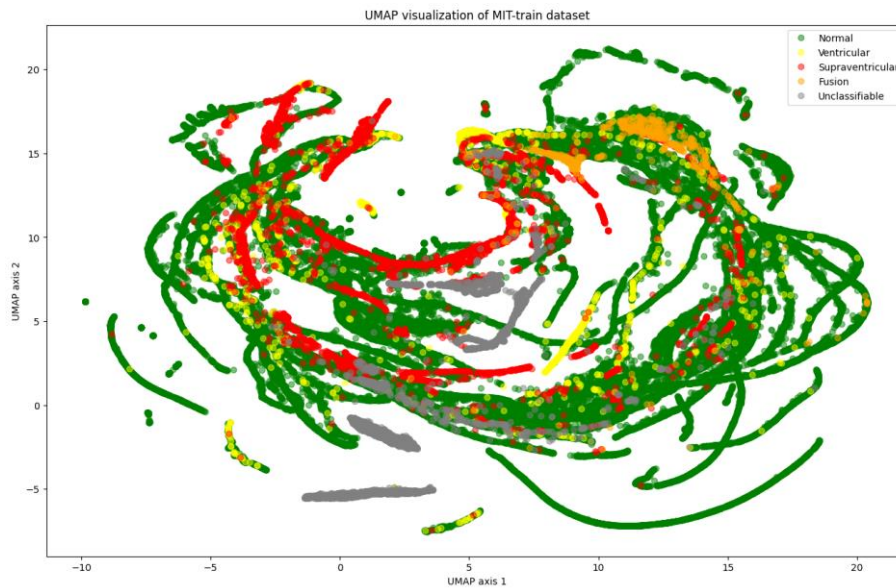2 place de Barcelona, 75016 Paris

Fig 8 - UMAP visualization showcasing the clustering of heartbeat types within the MIT-train dataset. Each cluster corresponds to a different heartbeat classification, illustrating both the diversity and similarity within heartbeat clusters.

Data rescaling (Normalization):

Rescaling, also referred to as normalization or standardization, is a common practice in machine learning preprocessing. It ensures that all features contribute equally to the model, preventing features with larger ranges from dominating the learning process. Hence, scaling can improve model performance, reduce sensitivity, and enhance training stability. In our analysis, by cross-validation, we aimed to evaluate the effectiveness of different rescaling methods to determine the best suitable scaler. The scalers that we considered were: StandardScaler, MinMaxScaler, and RobustScaler.

To address the performance of the mentioned scalers we conducted experiments on the dataset with each scaler and evaluated it by the average F1-score with different machine learning models. We applied multiple models —Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM)— using the differently scaled test data.

Table 1 - Mean F1-Scores for different combinations of rescaling methods and models

	StandardScaler	MinMaxScaler	RobustScaler	None
LogisticRegression	0.40925	0.40861	0.40904	0.40866
RandomForest	0.84545	0.84856	0.84703	0.84765
SVM	0.82775	0.82080	0.78088	0.8208

The results indicate that StandardScaler generally provides slightly better performance than MinMaxScaler and RobustScaler or no scaler. Using the robust scaler can obscure meaningful patterns in clean data by reducing sensitivity to variability and misaligning data centering, which is crucial for many machine learning algorithms. Similarly, the Min-Max scaler compresses all data into a fixed

DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

range, potentially amplifying the effect of outliers and misrepresenting feature variability. Both of these methods can lead to reduced model effectiveness, particularly in tasks requiring the detection of subtle differences, such as distinguishing between normal and abnormal heartbeats. Not scaling the data at all preserves its original distribution and variability, but this can cause issues with algorithms that assume normalized input. In contrast, the standard scaler preserves data variability and ensures proper centering, making it more suitable for clean datasets and enhancing model performance. Generally speaking, StandardScaler achieved the highest average F1-score than others. Although the performance difference across scalers is modest, StandardScaler is consistently effective. Therefore, we will use StandardScaler during preprocessing for future model training, given its slight edge in performance.

Data resampling:

Given the intrinsic imbalance in the dataset, which predominantly favors the normal heartbeat class, we addressed this challenge by employing several resampling techniques. The key methods included Synthetic Minority Over-sampling Technique (SMOTE), Random Oversampling, Random Undersampling, and Balanced Random Forest. SMOTE enhances the representation of the minority class by generating synthetic samples, while Random Oversampling duplicates existing minority class samples to balance the class distribution. Conversely, Random Undersampling mitigates imbalance by reducing the number of majority class samples. Balanced Random Forest addresses imbalance by adjusting class weights or employing resampling techniques during the training process to ensure a more balanced class distribution. Given the distribution, in our dataset of heartbeats with normal heartbeats being more common than ones it's crucial to rebalance our data. This way we prevent our model from favoring the group. It's also necessary to modify our loss function to address this discrepancy enabling learning from the class.

We evaluated these resampling methods in conjunction with a set of the previously mentioned machine learning models: LR, RF, SVM. To identify the optimal combination of model and resampling method, we employed Stratified K-Fold Cross-Validation. This technique ensures that each fold maintains the same class distribution as the original dataset, providing a reliable assessment of model performance. The models were evaluated using the mean F1-score, a metric well-suited for imbalanced datasets as it balances precision and recall. The mean F1 scores obtained for each sampling method and model are as follows:

Table 2 - Mean F1-Scores for different combinations of resampling methods and models.

	SMOTE	OverSampling	UnderSampling	Balanced Random Forest
LogisticRegression	0.407	0.410	0.405	0.409
RandomForest	0.877	0.866	0.810	0.848
SVM	0.829	0.828	0.778	0.821

Among the tested methods, LR consistently yielded relatively low scores across all resampling techniques, with the highest mean F1-score of 0.410 achieved using Random Oversampling. In contrast, SMOTE emerged as the most effective resampling technique, particularly when used in conjunction

DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

with RF, which achieved a superior mean F1-score of 0.877. The SVM model also performed well with SMOTE, achieving a mean F1-score of 0.829.

Overall SMOTE, especially when paired with RF, demonstrated the most substantial improvement in model performance, highlighting its effectiveness in addressing class imbalance in the dataset.

Conclusion:

In our first report, we prepared for the application of machine learning and deep learning techniques to detect and classify heartbeat normality and abnormality using ECG signals. Our study focused on data preparation, including exploration, preprocessing, feature engineering, and model setup for future applications.

We utilized two primary datasets—the MIT-BIH Arrhythmia Database and the PTB Diagnostic ECG Database—which provided a solid foundation for our preparations. Our preprocessing steps ensure data consistency and facilitate effective model training.

Exploratory data analysis, including visualizations and dimensionality reduction techniques such as PCA, t-SNE, and UMAP, revealed distinct patterns in ECG signals that enhanced our understanding of heartbeat classifications. These techniques helped us visualize the data structure and relationships between variables.

To benefit from the rescaling, we applied several rescaling methods. Finally, we concluded that StandardScaler is more suitable for our case. It can help provide a better f1-score compared to other scaler or no-scaler cases. It will help improve model performance, ensure fair feature contribution, and enhance the stability and speed of the training process.

To address the significant class imbalance, we applied various resampling techniques, with SMOTE demonstrating notable effectiveness. SMOTE generated synthetic samples for the minority class, which, when combined with models like RF, is expected to achieve the highest mean F1-score. This approach is crucial for improving diagnostic accuracy and enabling early detection of cardiac anomalies.

DataScientest.com
Training organization approval 11755665975
09 80 80 79 49
2 place de Barcelona, 75016 Paris

Bibliography:

1. Mark RG, Schluter PS, Moody GB, Devlin, PH, Chernoff, D. An annotated ECG database for evaluating arrhythmia detectors. *IEEE Transactions on Biomedical Engineering* 29(8):600 (1982).
2. Moody GB, Mark RG. The MIT-BIH Arrhythmia Database on CD-ROM and software for use with it. *Computers in Cardiology* 17:185-188 (1990).
3. Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. *IEEE Eng in Med and Biol* 20(3):45-50 (May-June 2001). (PMID: 11446209)
4. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation [Online].* 101 (23), pp. e215–e220.