Project Overview

We are collaborating with Deloitte on a machine learning project aimed at detecting cardiovascular diseases using ECG data. The objective is to develop a robust model that can classify different types of arrhythmias from ECG signals, aiding in early diagnosis and improving patient outcomes. This report summarizes the first prototype we have built, detailing the dataset, preprocessing, model architecture, training process, evaluation metrics, and results.

Dataset and Preprocessing

We used a labeled ECG dataset containing 21,892 records classified into 5 categories:

- Class 0: Normal
- Class 1: Arrhythmia Type 1
- Class 2: Arrhythmia Type 2
- Class 3: Arrhythmia Type 3
- Class 4: Arrhythmia Type 4

Key preprocessing steps included:

- Data balancing using resampling techniques to handle class imbalance.
- Signal normalization to scale the ECG readings between 0 and 1.
- Splitting data into training and validation sets (typically 80/20 split).
- One-hot encoding of categorical labels.

Model Architecture

We developed a neural network-based model using TensorFlow and Keras. The architecture consists of:

- Input Layer: Matching the length of ECG input data.
- Multiple 1D Convolutional Layers: To extract spatial features from the ECG waveforms.
- Batch Normalization and Dropout Layers: To prevent overfitting and speed up convergence.
- Dense Layers: Fully connected layers for final classification.
- Output Layer: Softmax activation for multi-class classification.

Training Details

- Optimizer: Adam
- Loss Function: Categorical Crossentropy
- Learning Rate: Default (0.001)
- Epochs: 20
- Batch Size: 32
- Callbacks: EarlyStopping (to prevent overfitting), ModelCheckpoint (to save the best model)

Evaluation Metrics and Results

The model achieved the following classification report metrics:

Class	Precision	Recall	F1-Score	Support
0	1.00	0.97	0.98	18118
1	0.56	0.87	0.68	556
2	0.93	0.96	0.94	1448
3	0.60	0.85	0.71	162
4	0.95	0.99	0.97	1608

- **Overall Accuracy:** 97%
- **Macro Average F1-Score:** 0.86
- **Weighted Average F1-Score:** 0.97

Confusion Matrix Insights

The confusion matrix shows that the model performs very well for Class 0 and Class 4, with minimal misclassifications. However, it struggles more with Class 1 and Class 3, which have lower precision and F1-scores, likely due to smaller sample sizes and overlapping ECG features between classes.

Current Limitations

- The model currently includes only 4 types of arrhythmias (Classes 1-4) besides the normal class.
- Class imbalance remains a challenge, especially for Classes 1 and 3, which are underrepresented.

- Only basic CNN architecture is used; no advanced techniques like transfer learning or attention mechanisms are yet implemented.

Further Improvements

For the next iteration of the project, we plan the following improvements:

- **Inclusion of More Arrhythmia Types:** We aim to expand the dataset to include a broader range of arrhythmias, making the model more comprehensive.
- **Advanced Data Augmentation:** Implementing synthetic data generation (e.g., using GANs or SMOTE for time-series data) to address class imbalance more effectively.
- **Model Architecture Enhancements:** Introducing more advanced architectures such as Residual Networks (ResNet) or Transformers specifically designed for time-series data.
- **Feature Engineering:** Extracting additional features like heart rate variability (HRV), PQRST interval measurements, and frequency-domain features.
- **Hyperparameter Tuning:** Using Grid Search or Bayesian Optimization to fine-tune hyperparameters for better performance.
- **Cross-Validation:** Implementing k-fold cross-validation to ensure the model's robustness and generalization capability.
- **Explainability:** Adding explainability modules like Grad-CAM or SHAP to make model predictions interpretable to medical professionals.
- **Real-time Inference:** Exploring optimization techniques to make the model lightweight and capable of running in real-time on edge devices.