ISyE 6740 – Spring 2021
Project Proposal

Team Member Names: Eric Goldberg, Ojas Mehta, Phumthep Bunnak

Project Title: Exploring the Application of Supervised Learning for Cardiovascular Disease Prediction

1 Background

Electrocardiography (ECG) is a common, non-invasive diagnostic procedure usually conducted and interpreted by clinicians. The ECG provides information regarding the electrical activity and conduction through cardiac tissue. This information has traditionally been utilised to provide judgement on pathology. Although the ECG has a significant amount of information, it is often used along with clinical information and other investigations to make a diagnosis. Some common diagnoses are Acute Myocardial Infarction, Left Ventricular Hypertrophy and Cardiac Arrhythmias.

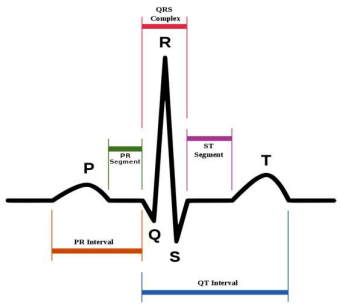

Figure 1: ECG example from Wikipedia

As data analytics gains popularity in many industries from finance to insurance, we expect ECG interpretation to become increasingly automated in the future. This is especially important considering rising demand for telemedicine (Koonin et. al, 2020) and the potential of data from "wearables" such as smartwatches (Isakadze & Martin, 2020). Doctors and medical systems will need tools to cope with these expanding volumes of data especially given the inconsistencies that exist between clinicians across a range of medical specialties (Salerno, Alguire, & Waxman, 2003). The main obstacle to develop a machine learning model for electrocardiography was the lack of high quality, labelled datasets created from digitized ECGs. Since, many ECG datasets have been made available both for free and under commercial licenses over the past few years[1]. Accessible ECG datasets for model learning and the widespread adoption of sensor-laden wearables such as smartwatches will enable the application of real-time monitoring and diagnosing health conditions.

A few research groups have published promising results on the application of deep learning models to ECG interpretation. For examples, see Strodthoff et al. (2020) as well as Hannun et al (2019).

2 Problem Statement

Our primary objective is to develop and optimize a machine learning model utilizing ECG and patient data to predict tier 1 cardiac diagnoses[2] using a labelled dataset. We will use the following supervised learning models: SVM, logistic regression, naïve bayes, k-nearest neighbors, MLP neural network, and decision trees.

[1] Examples include the MIT-BIH Arrythmia Database and iRhythm technologies.
[2] See Table 1 in Data Source for tier 1 diagnoses descriptions.

Additional secondary objectives include the following:

A. Predict tier 2 diagnoses given a tier 1 diagnosis.
B. Determine the effect of data reduction (e.g., linear and non-linear dimensionality reduction of ECG data with PCA) on model performance e.g., use of select leads rather than entire 12 lead ECG information.
C. The effect of parameter tuning on model performance with active inclusion and exclusion of clinical data and patient factors e.g., age, weight, height, gender.
D. Can we process (detrend, first order differences, etc.) the data to improve model performance?

3 Data Source

This project uses one of the largest clinical ECG datasets called PTB-XL, made available by researchers at the Physikalisch Technische Bundesanstalt (PTB) (Wagner et al., 2020). The data was collected from 18,885 patients between 1989 to 1996 using the Schiller AG ECG device. The dataset contains 21,837 clinical 12-lead ECG records of 10 seconds length. This digitized ECG data comes in the form of a numerical matrix with floats representing deviations from the baseline. The researchers at PTB designed the dataset to be balanced with respect to sex and to cover a wide range of ages.

The ECG was interpreted and annotated by cardiologists whose statements are in accordance with the SCP-ECG (Standard communications protocol for computer assisted electrocardiography). The SCP statement descriptions were grouped into diagnostic classes below, also see Figure 2 in the appendix:

Table 1: SCP-statements

Tier 1 Diagnoses		Tier 2 Diagnoses
CD	Cardiac conduction disturbances	• Wolff-Parkinson-White or Incomplete Left Bundle Branch Block (WPW, ILBBB) • Complete Left Bundle Branch Block (CLBBB) • Complete Right Bundle Branch Block (CRBBB) • Intraventricular Conduction Delay (IVCD) • Atrioventricular Block (_AVB) • Incomplete right bundle branch block (IRBBB) • Left Anterior Fascicular Block / Left Posterior Fascicular Block (LAFB/ LPFB)
HYP	Ventricular Hypertrophy	• Right Ventricular Hypertrophy (RVH) • Right Atrial overload/enlargement (RAO/RAE) • Left Atrial overload/enlargement (LAO/LAE) • Septal hypertrophy (SEHYP) • Left Ventricular Hypertrophy (LVH)
MI	Myocardial infarction	• Lateral Myocardial Infarction (LMI) • Posterior Myocardial Infarction (PMI) • Anteroseptal Myocardial Infarction (AMI) • Inferior Myocardial Infarction (IMI)
NORM	Normal ECG	• Normal ECG
STTC	ST-T descriptive statements	• ST-T change (STTC) • Ischemic ST-T change (ISC_)/ in anterior leads (ISCA)/ in inferior leads (ISCI) • Non-specific ST changes (NST_)

For further details on the data acquisition and the preprocessing steps, please refer to Wagner et al (2020).

4 Methodology

4.1 Data preprocessing step (in progress and will appear in the final report)

We based our preprocessing step on the example Python script from Wagner et al (2020) for data preprocessing. The purpose of the script was to load waveform and corresponding meta-data (id, age, sex, diagnostic class, etc.) into an array and dataframe. The main outputs from the preprocessing script include an array, X, of shape (21837, 1000, 12) containing the 12-lead ECG signal for 21837 samples, and a dataframe Y, containing patient meta-data for each ECG such as id, age, sex, recording date, recording device, and the diagnostic classes. In this section, we briefly describe how the preprocessing step works.

The first step was to import the ECG data which is kept in the WaveForm DataBase (WFDB) format. We used the "WFDB" python package to read in the downsized time series with a sampling rate of 100 Hz. This means the resulting ECG dataset was a time series with 1000 points for 10 seconds length per lead. Afterwards, we plan to extract morphological features from the ECGs such as R-R intervals, Q-R-S width, P waves, P-R segment and S-T segment (Bazi et al. 2013).

Following Wagner et al., we preserved the training (90%) and testing (10%) set split used by the authors in order to have comparability in model test performance with others using the same data set. The splitting process also stratified the data such that the training set and the test set contain similar shares of diagnostic classes. This ensured balanced diagnostic results in both the training set and the test set. Folds 1-8 are intended for training, with fold 9 for validation, and fold 10 for testing.

4.2 Fitting models and hyperparameter tuning

We will treat the dataset as a multi-label classification problem using "binary relevance." Binary relevance decomposes the multi-label learning problem into q independent binary learning problems (Zhang et. al, 2018). This project will explore the application of linear and nonlinear classification models: support-vector machine, logistic regression, Naïve Bayes, K-nearest neighbors, and MLP neural network. Details are provided in the table 2 in the appendix.

Most of our models require careful selection of hyperparameters. We will use grid search or random grid search coupled with 9-fold cross validation for hyper parameter tuning. For example, the K-nearest neighbor algorithm requires selecting the optimal number of neighbors.

5 Evaluation and Final Results (in progress and will appear in the final report)

Since our project concerns multi-label classification, we need to adopt a different set of evaluation criteria from that of a binary classification problem. In binary classification with given labels, the predictions could be grouped into correct predictions and incorrect predictions. However, in a multi-label classification problem, we would give some value or positive score to "half-right" predictions, which contain some of the correct labels. Nooney (2018) suggested using micro and macro-averaging for this purpose:

5.1 Evaluating Model Performance

5.1.1 Micro-averaging

For each class, the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are summed up and averaged. Micro-averaging is preferred when the classes are not distributed proportionally[3].

[3] Micro Average vs Macro average Performance in a Multiclass classification setting. (2017). Retrieved March 31, 2021, from https://datascience.stackexchange.com/questions/15989/micro-average-vs-macro-average-performance-in-a-multiclass-classification-settin

- Micro-average precision

$$P_{c_i} = \frac{\sum_{c_i \in C} TP_{c_i}}{\sum_{c_i \in C} TP_{c_i} + FP_{c_i}}$$

for C = diagnostic classes

- Micro-averaging recall

$$R_{c_i} = \frac{\sum_{c_i \in C} TP_{c_i}}{\sum_{c_i \in C} TP_{c_i} + FP_{c_i}}$$

5.1.2 Macro-average

The precision and recall for different classes are calculated first and then the average is calculated. This means we are weighting the metrics of each class equally.

- Macro-average precision

$$P = \frac{\sum_{c_i \in C} P_{c_i}}{|k|}$$

- Macro-average recall

$$R = \frac{\sum_{c_i \in C} R_{c_i}}{|k|}$$

For k = number of diagnostic classes

5.1.3 Hamming Loss

Alternatively, we might consider evaluating the models using the Hamming loss which is the fraction of incorrectly predicted labels. Hamming loss represents the Hamming distance between the predicted label and true label. The Sci-kit learn package calculates Hamming loss as:[4]

$$L_{Hamming}(y, \hat{y}) = \frac{1}{n_{labels}} \sum_{j=0}^{n_{labels}-1} 1(\hat{y} \neq y)$$

For $1(x)$ = indicator function

5.2.1 Assessing the trade-off between reductions in ECG data and prediction accuracy

After fitting the 12-lead ECG signal to different models and the best performing model is selected, we will use the resulting model to determine the minimum number of leads without sacrificing too much accuracy. Instead of comparing different models, we will compare the performance of the same model but fitted with different numbers of feature. Ideally, we should be able to identify the most important feature, which if excluded would greatly reduce the model's performance. We will also determine the effect of dimensionality reduction on model performance. We expect that model computing time will improve with dimensionality reduction using tools such as principal component analysis.

[4] Metrics and scoring: quantifying the quality of predictions. Retrieved March 31, 2021, from https://scikit-learn.org/stable/modules/model_evaluation.html#hamming-loss

5.3 Final Results (in progress and will appear in the final report)

6 Summary and Next Steps

Above, we have described our key objectives, data source, proposed methodology, and evaluation. Our next immediate steps are processing our data and building the models. We anticipate our main challenges to be in feature extraction from the wfdb files and in parameter tuning. Moreover, almost all other work reviewed has used deep learning. We hope to accomplish comparable results but may not reach the same performance levels. However, we believe that our examination of patient meta-data inclusion and reduced lead data contributes to the existing literature.

7 Citations

1. Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, *25*(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3

2. Isakadze, N., & Martin, S. S. (2020). How useful is the smartwatch ECG? *Trends in Cardiovascular Medicine*, *30*(7), 442–448. https://doi.org/10.1016/j.tcm.2019.10.010

3. Koonin, L. M., Hoots, B., Tsang, C., Leroy, Z., Farris, K., & Jolly, B. T. (2020). Trends in the Use of Telehealth During the Emergence of the COVID-19 Pandemic — United States, January–March 2020. *Morbidity and Mortality Weekly Report (MMWR)*, *69*(43), 1595–1599. https://www.cdc.gov/mmwr/volumes/69/wr/mm6943a3.htm?s_cid=mm6943a3_w

4. Nooney, K. (2019, February 12). *Deep dive into multi-label classification..! (With detailed Case Study)*. Medium. https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff

5. Salerno, S. M., Alguire, P. C., & Waxman, H. S. (2003). Competency in Interpretation of 12-Lead Electrocardiograms: A Summary and Appraisal of Published Evidence. *Annals of Internal Medicine*, *138*(9), 751. https://doi.org/10.7326/0003-4819-138-9-200305060-00013

6. Strodthoff, N., Wagner, P., Schaeffter, T., & Samek, W. (2020). Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 1. https://doi.org/10.1109/jbhi.2020.3022989

7. Wagner, P., Strodthoff, N., Bousseljot, R. D., Kreiseler, D., Lunze, F. I., Samek, W., & Schaeffter, T. (2020). PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, *7*(1). https://doi.org/10.1038/s41597-020-0495-6

8. Zhang, M. L., Li, Y. K., Liu, X. Y., & Geng, X. (2018). Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, *12*(2), 191–202. https://doi.org/10.1007/s11704-017-7031-7

8 Appendix

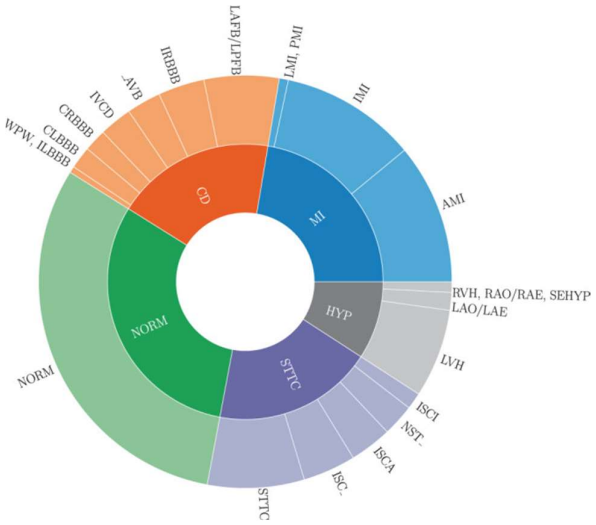

Figure 2: Graphical summary of the PTB-XL dataset in terms of diagnostic superclasses and subclasses from Wagner et al (2020)

Table 2: Model Summary and Parameters

Model	Model Description and applicability to ECG application	Parameters*
Support-Vector Machines (SVMs)	• Package: sklearn.svm.SVC • SVMs were shown to have a high-level of accuracy (90%+) in predicting cardiac arrhythmia from ECG signals.[5,6]	"C": the regularization parameter "kernel": kernel type for the algorithm "gamma": coefficient for the kernel
Logistic regression	• Package: sklearn.linear_model.LogisticRegression • Rahman et al (2015) applied random forest classifier, a SVM, and logistic regression to identify hypertrophic cardiomyopathy using 12-lead ECG signals.[7] Logistic regression was outperformed by the two models.	"C": the regularization parameter
Naïve Bayes	• Package: sklearn.naive_bayes.GaussianNB • Padmavathi and Ramanujam (2015) applied Naïve Bayes model to classify abnormal heart condition using ECGs from the MIT–BIH Arrhythmia database.[8]The result showed an accuracy of 93.3%.	-
K-nearest neighbors	• Package: sklearn.neighbors.KNeighborsClassifier	"num_neighbors": the number of neighbors

[5] Polat, K., & Güneş, S. (2007). Detection of ECG Arrhythmia using a differential expert system approach based on principal component analysis and least square support vector machine. *Applied Mathematics and Computation*, *186*(1), 898-906.
[6] Nasiri, J. A., Naghibzadeh, M., Yazdi, H. S., & Naghibzadeh, B. (2009, November). ECG arrhythmia classification with support vector machines and genetic algorithm. In 2009 Third UKSim European Symposium on Computer Modeling and Simulation (pp. 187-192). IEEE.
[7] Q. A. Rahman, L. G. Tereshchenko, M. Kongkatong, T. Abraham, M. R. Abraham and H. Shatkay, "Utilizing ECG-Based Heartbeat Classification for Hypertrophic Cardiomyopathy Identification," in IEEE Transactions on NanoBioscience, vol. 14, no. 5, pp. 505-512, July 2015, doi: 10.1109/TNB.2015.2426213.
[8] Padmavathi, S., & Ramanujam, E. (2015). Naïve Bayes classifier for ECG abnormalities using multivariate maximal time series motif. *Procedia Computer Science*, *47*, 222-228.

	• KNN showed an accuracy of 98.71% in classifying ECGs from the MIT-BIH arrhythmia database into categories: categories of beats are: Normal (N), Premature Ventricular Contraction (PVC), Atrial Premature Contraction (APC), Right Bundle Branch Block (RBBB) and Left Bundle Branch Block (LBBB).[9] Another version of the KNN called evidential k nearest neighbours (EKNN) approach which is based on Dempster Shafer Theory could defer classification to avoid incorrectly classifying the ECG sinal.[10]	
Classification and Regression Tree	• Package: sklearn.tree.DecisionTreeClassifier • A group of researchers fitted 1-lead ECG signal to the random forest model to classify the ECG signal into normal, suspicious to AF, suspicious to other arrhythmia, noise.[11] Their result showed an F-1 score of 0.81.	"max_depth": the depth of the tree "min_samples_leaf" and "max _leaf_nodes": the minimum/maximmum number of samples required as leaf nodes "max_features": the number of features to determine the best split
MLP neural network	• Package: sklearn.neural_network.MLPClassifier • Different types of neural network models for classifying abnormal heart conditions based on ECG signal were studied such as Quantum Neural Networks, Robust Neural Network-Based Classification, and Block-Based Neural Networks.[12]	"hidden_layer_sizes" "solver": the solver for weight optimization "alpha": the regularization parameter "learning rate": the initial learning rate

**based on parameters from the Sci-Kit Learn package.*

[9] F. Bouaziz, D. Boutana and H. Oulhadj, "Diagnostic of ECG Arrhythmia using Wavelet Analysis and K-Nearest Neighbor Algorithm," 2018 International Conference on Applied Smart Systems (ICASS), Medea, Algeria, 2018, pp. 1-6, doi: 10.1109/ICASS.2018.8652020.

[10] Faziludeen, S., & Sankaran, P. (2016). ECG beat classification using evidential K-nearest neighbours. Procedia Computer Science, 89, 499-505.

[11] M. Kropf, D. Hayn and G. Schreier, "ECG classification based on time and frequency domain features using random forests," 2017 Computing in Cardiology (CinC), Rennes, France, 2017, pp. 1-4, doi: 10.22489/CinC.2017.168-168.

[12] Padmavathi, S., & Ramanujam, E. (2015). Naïve Bayes classifier for ECG abnormalities using multivariate maximal time series motif. *Procedia Computer Science*, *47*, 222-228.