A Machine Learning Approach for ECG Classification on the PTB-XL Dataset

Tirtho Das
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh

Nura Alam
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh

I. Introduction

Electrocardiography (ECG) is a popular non-invasive method for the analysis and classification of cardiovascular disorders. The proper classification of ECG signals is very important for efficient identification of cardiovascular disorders including conduction disorders (CD), hypertension-related patterns (HYP), myocardial infarction (MI), normal rhythms (NORM), and ST-T changes (STTC). The manual analysis of ECG signals is a time-consuming and subjective process that sometimes results in different conclusions between observers, thereby pointing towards the need for automated classification techniques.

Deep learning techniques like CNN have demonstrated large potentialities for the extraction of temporal andmorphological features from the ECG signal. Supervised deep learning techniques like the InceptionTime model and the ResNet model require a large number of labeled examples for high accuracy classification. On the other hand, the MobileNetV2 model is light enough to run on resource-constrained settings. However, the reliance on a large number of labeled examples is a major drawback of the former.

Self-supervised learning handles this limitation by learning informative descriptors from unsupervised data, which can then be utilized for downstream classification learning. Recently, approaches like MAE (Masked AutoEncoders), TS-JEPA, MoCo v2, and BYOL have shown their capability of learning meaningful patterns of signal, making it feasible to learn ECG representations with very few labelled observations.

This analysis examines the classification ability of both supervised and self-supervised methods on a publicly available ECG dataset to compare their ability to handle classification tasks well and adaptability to different categories. The aim is to find suitable ways of performing ECG classification.

II. Related Work

Recently, deep learning has become a promising alternative to traditional methods like rule-based and feature-based machine learning for classifying electrocardiograms. Traditional techniques often depend on extensive human analysis and manually designed features. In contrast, CNN has become popular for processing vital signs due to its capability to recognize intricate patterns within high-dimensional signals.

A. Architectural Developments & Benchmarking

Standardization and benchmarking have played a vital role in advancing research in this area. To establish benchmarks on the PTB-XL dataset, Strodthoff et al. evaluated ResNet and inception-based architectures, discovering that these models performed optimally. They also investigated various neural network architectures. Their xresnet1d101 network achieved a macro AUC of 0.96 for rhythm classification and an accuracy of 89.8% for sex prediction tasks. To address data challenges such as class imbalance, Khan et al. created a 1-D convolutional deep Residual Neural Network known as ResNet, utilizing the Synthetic Minority Over-sampling Technique (SMOTE) to achieve a mean accuracy of 98.63%. Hybrid approaches have also been explored to consider the temporal dimensions. Cheng et al. developed a 24-layer DCNN integrated with Bidirectional Long Short-Term Memory (BiLSTM) and a Tan Mean Square Error (TMSE) loss function, achieving an accuracy of 89.3% in their predictions.

B. Transfer Learning and Multi-Lead Fusion

Transfer learning has proven essential due to the limited availability of labeled medical data. Rahman et al. introduced the CAA-TL model that utilized pre-trained AlexNet (98.38% accuracy), ResNet50 (91% accuracy), and SqueezeNet (90.08% accuracy) for classifying five types of arrhythmia. Weimann and Conrad (2021) demonstrated that pre-training CNNs with the Icentia11K dataset significantly enhanced AF classification results by as much as 6.57%. Furthermore, they investigated the spatial relationships among ECG leads to improve AF classification models. Park et al. suggested a data fusion technique where the model learned from various individual single-lead ECGs, concluding that the optimal diagnostic leads are Leads aVR and II, achieving F1 scores of 98.7% and 98.2%, respectively.

C. Paradigms of Self-Supervised

Recent studies have begun to focus on Self-Supervised Learning (SSL), which aims to utilize unlabeled data. Soltanieh et al. explored Self-Supervised Learning techniques like SimCLR, BYOL, and SwAV, and found that SwAV could acquire effective representations, obtaining the highest Macro F1 score. Backed primarily by data augmentation, Atienza et al. introduced SBnCL, a contrast-free approach based on

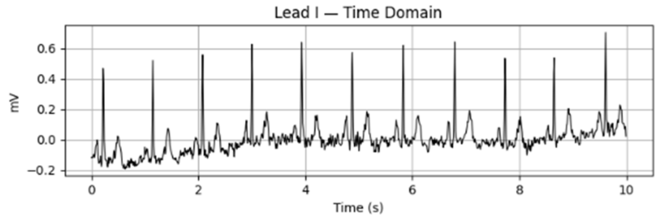

Fig. 1. ECG signal visualization

Vision Transformers that does not rely on data augmentation or negative samples, reaching 95.7% accuracy for gender classification. Weimann and Conrad (2023) proposed Joint Embedding Predictive Architecture (JEPA), which focuses on predicting representations instead of reconstructing inputs, achieving a leading AUC of 0.945 on the PTB-XL datasets. In a related initiative to address the spatial correlations among leads in a multilead ECG signal, Chen et al. developed Temporal-Spatial Self-Supervised Learning (TSSL). This method merges self-supervision techniques of Temporal Self-Supervision and Spatial Self-Supervision to preserve inter-lead relationships, achieving a macro AUC of 0.882 on the PTB-XL dataset using only 10% labeled data, approaching the results of full labeling.

III. Methodology

A. Dataset and Preprocessing

We used the PTB-XL ECG dataset, which contains 21,799 12-lead ECG records of 10-second duration. After filtering for single-label diagnostic classes, we retained 16,244 records across five main categories: Normal (NORM), Myocardial Infarction (MI), ST/T Change (STTC), Conduction Disturbance (CD), and Hypertrophy (HYP). All signals were resampled to 100 Hz, truncated or zero-padded to a fixed length of 1,000 samples, and normalized, resulting in a consistent input shape of (12, 1000) suitable for 1D CNN-based models.

B. Data Splitting

The experimental framework utilized a sliding scale of data partitions for model development and evaluation. These allocations ranged from a training-dominant 9:1 ratio to a testing-dominant 1:9 ratio, encompassing all decile increments (80-20, 70-30, 60-40, 50-50, 40-60, 30-70, and 20-80)

C. Supervised Learning Models

ResNet-1D : A 1D adaptation of ResNet-18 using residual connections to mitigate vanishing gradients. This model achieved the best overall supervised performance, especially for dominant classes.
MobileNetV2-1D : A lightweight architecture using depthwise separable 1D convolutions, optimized for efficiency.
Inception Time : InceptionTime employs parallel convolutions with multiple kernel sizes to capture ECG patterns at different temporal scales.

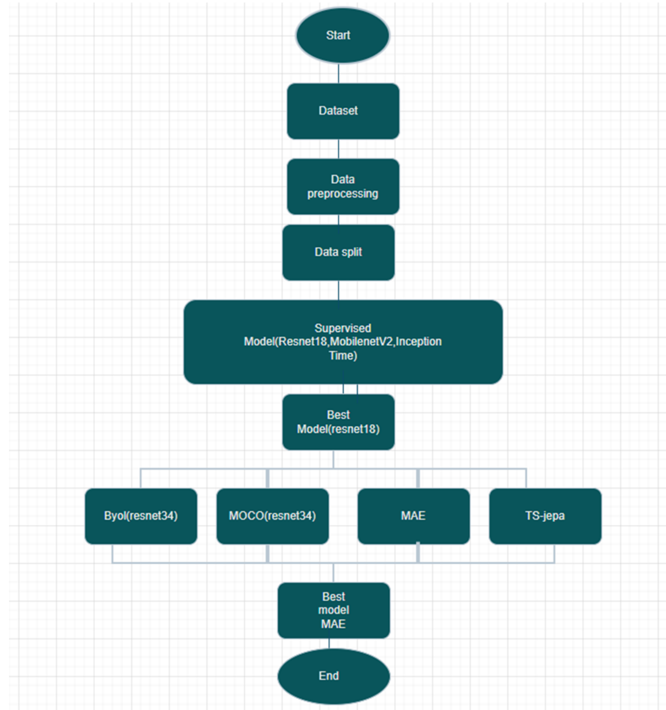

Fig. 2. Workflow diagram

D. Self-Supervised Learning

We conducted a series of self-supervised experiments on the PTB-XL 12-lead ECG dataset to evaluate different representation learning methods. The dataset was preprocessed by resampling all signals to 100 Hz, truncating or zero-padding them to a fixed length of 1,000 samples, and normalizing to ensure consistent input shape for 1D CNN models. The first approach, MoCo v2 (Momentum Contrast), employed a ResNet-1D backbone to generate 256-dimensional embeddings for each ECG signal. Contrastive learning was used with a dynamic memory queue, where augmented views of the same signal were pulled closer in the embedding space while embeddings of other signals were pushed apart, minimizing the NT-Xent loss. Data augmentations included jittering, scaling, masking of random time segments, and channel dropout. MoCo v2 was trained for 50 epochs using the Adam optimizer with a learning rate of 1e-3.For downstream classification, embeddings were extracted from the frozen encoder and fed into Logistic Regression and SVM classifiers,Random ForestDT. The second method, BYOL (Bootstrap Your Own Latent), also used a ResNet-1D backbone. Unlike contrastive approaches, BYOL predicts the representation of a target network—maintained as an EMA of the online network—from the online network output, and therefore does not require negative pairs. Augmentations were identical to those used in MoCo v2, and training ran for 50 epochs with momentum updates for the target networkFor downstream classification, embeddings were extracted from the frozen encoder and fed into Logistic Regression and SVM classifiers,Random

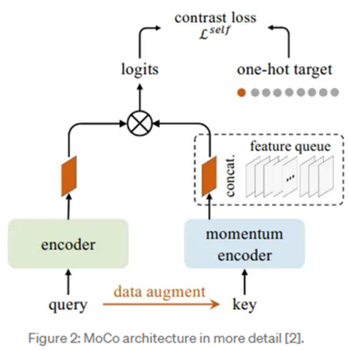

Fig. 3. MOCO architecture

Forest,DT. The third method, MAE (Masked Autoencoder),

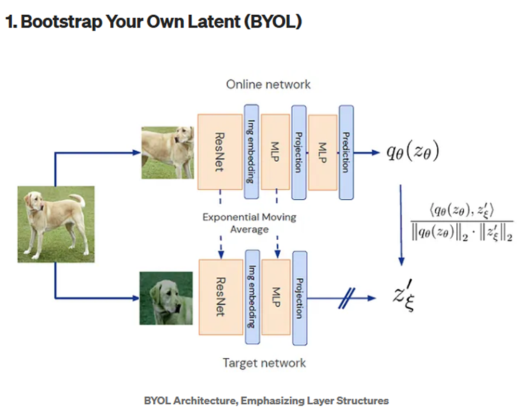

Fig. 4. BYOL

used a ResNet-1D encoder with a decoder reconstruction head. Randomly masking 25% of the signal segments, the model was trained to reconstruct the masked parts using mean squared error loss. Positional encoding was added to help capture temporal structure, and training was performed for 100 epochs with AdamW optimizer and gradient clipping.For downstream classification, embeddings were extracted from the frozen encoder and fed into Logistic Regression and SVM classifiers,Random Forest,MLP,DT.

Finally, TS-JEPA (Temporal-Spatial Joint Embedding Predictive Architecture) represented ECG signals as patches processed through a ResNet-1D backbone, augmented with jittering, scaling, random shifts, and time masking. Patch embeddings were fed into a transformer-based context network that captured temporal relationships, and a target network maintained via EMA provided stable predictions for masked patches. The model minimized the cosine similarity loss between predicted and target embeddings, with the online network updated via backpropagation and the target network via EMA. For downstream classification, embeddings were

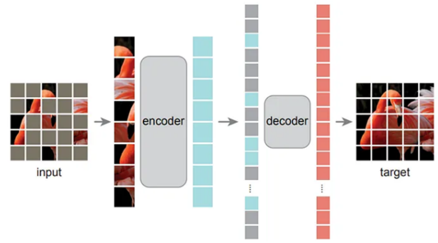

Fig. 5. MAE architecture

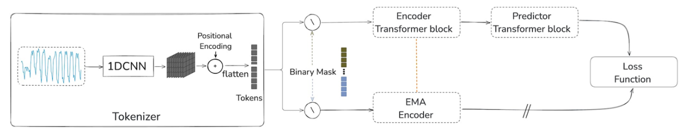

Fig. 6. TS-JEPA architecture

extracted from the frozen encoder and fed into Logistic Regression and SVM classifiers.

IV. Result and Discussion

The results of these supervised learning models were tested on a large publicly available ECG dataset that contains five classes: CD, HYP, MI, NORM, and STTC. The results of InceptionTime showed an overall accuracy of 0.643 to 0.757 on various split ratios of the dataset, and the best accuracy was found at a split ratio of 70:30. The highest precision and F1 score were found in class NORM, and the score reached a maximum of 0.907 and 0.825, respectively, showing that it is easier to identify normal ECG signals. On the contrary, it was most difficult to identify class HYP with an F1 score of less than 0.36, showing that it is quite tough to identify the subtle hypertensive pattern.

TABLE I
INCEPTIONTIME PERFORMANCE ON ECG DATASET ACROSS TRAIN–TEST SPLITS

Split	Acc.	Class	Prec.	Rec.	F1
10:90	0.686596	CD	0.682507	0.7254766	0.703336
		HYP	0.203187	0.714953	0.316443
		MI	0.644024	0.486673	0.554400
		NORM	0.862545	0.732497	0.792219
		STTC	0.690104	0.565032	0.621336
20:80	0.642955	CD	0.727205	0.633051	0.676870
		HYP	0.217854	0.551867	0.312390
		MI	0.882392	0.652659	0.750335
		NORM	0.455954	0.735648	0.562976
		STTC	0.882392	0.652659	0.750335
30:70	0.685076	CD	0.604326	0.794314	0.686416
		HYP	0.212568	0.730667	0.329327
		MI	0.577320	0.632054	0.603448
		NORM	0.902156	0.685570	0.779091
		STTC	0.626575	0.651190	0.638646
40:60	0.715913	CD	0.623267	0.789268	0.696513
		HYP	0.229853	0.781931	0.355272
		MI	0.623274	0.624095	0.623684
		NORM	0.884689	0.761301	0.818370
		STTC	0.717259	0.574306	0.637871
50:50	0.739350	CD	0.668072	0.742389	0.703272
		HYP	0.300325	0.690299	0.418552
		MI	0.698182	0.45497	0.550933
		NORM	0.831119	0.855315	0.843043
		STTC	0.686679	0.610000	0.646072
60:40	0.727608	CD	0.694696	0.786237	0.737637
		HYP	0.260870	0.728972	0.384236
		MI	0.610460	0.668312	0.638077
		NORM	0.900167	0.743109	0.814133
		STTC	0.647116	0.689583	0.667675
70:30	0.757078	CD	0.821586	0.728516	0.772257
		HYP	0.318750	0.633540	0.424116
		MI	0.682471	0.625000	0.652473
		NORM	0.890857	0.794928	0.840163
		STTC	0.591189	0.801389	0.680425
80:20	0.739304	CD	0.702290	0.807018	0.751020
		HYP	0.250696	0.543478	0.386266
		MI	0.743243	0.774531	0.627854
		NORM	0.894335	0.741667	0.830133
		STTC	0.6402	0.774531	0.687259
90:10	0.747077	CD	0.761364	0.783626	0.772334
		HYP	0.246835	0.722222	0.367925
		MI	0.705882	0.664032	0.684318
		NORM	0.907407	0.756340	0.825015
		STTC	0.629630	0.779167	0.696462

TABLE II
MOBILENETV2 PERFORMANCE ON ECG DATASET ACROSS TRAIN–TEST SPLITS

Split	Acc.	Class	Prec.	Rec.	F1
10:90	0.366347	CD	0.623188	0.754386	0.682540
		HYP	0.761364	0.783626	0.772334
		MI	0.246835	0.722222	0.367925
		NORM	0.863388	0.771202	0.696803
		STTC	0.203187	0.714953	0.316443
20:80	0.559403	CD	0.644024	0.486673	0.554400
		HYP	0.862545	0.732497	0.792219
		MI	0.623188	0.754386	0.682540
		NORM	0.863388	0.771202	0.696803
		STTC	0.577320	0.632054	0.603448
30:70	0.546654	CD	0.902156	0.685570	0.779091
		HYP	0.623188	0.754386	0.682540
		MI	0.882392	0.652659	0.750335
		NORM	0.863388	0.771202	0.696803
		STTC	0.882392	0.652659	0.750335
40:60	0.633426	CD	0.623188	0.754386	0.682540
		HYP	0.882392	0.652659	0.750335
		MI	0.455954	0.735648	0.562976
		NORM	0.623188	0.754386	0.682540
		STTC	0.623188	0.754386	0.682540
50:50	0.588402	CD	0.863388	0.771202	0.696803
		HYP	0.391905	0.469545	0.391905
		MI	0.623188	0.754386	0.682540
		NORM	0.902156	0.685570	0.779091
		STTC	0.863388	0.771202	0.696803
60:40	0.656664	CD	0.882392	0.652659	0.750335
		HYP	0.455954	0.735648	0.562976
		MI	0.623188	0.754386	0.682540
		NORM	0.623188	0.754386	0.682540
		STTC	0.902156	0.685570	0.779091
70:30	0.631514	CD	0.626575	0.651190	0.638646
		HYP	0.863388	0.771202	0.696803
		MI	0.577320	0.632054	0.603448
		NORM	0.623188	0.754386	0.682540
		STTC	0.668072	0.742389	0.703272
80:20	0.680209	CD	0.300325	0.690299	0.418552
		HYP	0.698182	0.45497	0.550933
		MI	0.831119	0.855315	0.843043
		NORM	0.623188	0.754386	0.682540
		STTC	0.761364	0.783626	0.772334
90:10	0.660923	CD	0.246835	0.722222	0.367925
		HYP	0.705882	0.664032	0.684318
		MI	0.907407	0.756340	0.825015
		NORM	0.623188	0.754386	0.682540
		STTC	0.633375	0.63337	0.679783

In the MobileNetV2 model, a light-weight convolutional neural network, there was a relatively lower average accuracy range of 0.366 to 0.680. Although it had a decent level of precision and F1 score values in the NORM and CD classes with around 0.77, rare classes like MI and STTC were not accurately distinguished with F1 scores in many instances below 0.38. Moreover, contrary to the expected trend in accuracy with increased training data, it was noticed that in a few instances, the accuracy scores were not necessarily enhanced with the rise in training datasets; this might be a consequence of the inherent capability of MobileNetV2 to comprehend intricate patterns in a limited model capacity scenario.

Results showed that ResNet18 was quite robust, with accuracy scores varying from 0.713 to 0.792, comparable to or slightly higher than those from InceptionTime. Both NORM and STTC classes showed high F1-score values above 0.79, though HYP detection remained difficult. Accuracy tended to be higher with higher proportions in training sets, peaking at 80:20 split levels. This is due to its ability to preserve the flow of gradients from activations via its connectionist architecture and extract hierarchical features. This paper acknowledges that there is a trade-off concerning complexity and accuracy with respect to model depth in ResNet18.

TABLE III
RESNET18 PERFORMANCE ON ECG DATASET ACROSS TRAIN–TEST SPLITS

Split	Acc.	Class	Prec.	Rec.	F1
10:90	0.712996	CD	0.668072	0.742389	0.703272
		HYP	0.300325	0.690299	0.418552
		MI	0.698182	0.45497	0.550933
		NORM	0.831119	0.855315	0.843043
		STTC	0.623188	0.754386	0.682540
20:80	0.729224	CD	0.882392	0.652659	0.750335
		HYP	0.455954	0.735648	0.562976
		MI	0.623188	0.754386	0.682540
		NORM	0.577320	0.632054	0.603448
		STTC	0.902156	0.685570	0.779091
30:70	0.741799	CD	0.626575	0.651190	0.638646
		HYP	0.623188	0.754386	0.682540
		MI	0.203187	0.714953	0.316443
		NORM	0.644024	0.486673	0.554400
		STTC	0.862545	0.732497	0.792219
40:60	0.752129	CD	0.668072	0.742389	0.703272
		HYP	0.300325	0.690299	0.418552
		MI	0.698182	0.45497	0.550933
		NORM	0.831119	0.855315	0.843043
		STTC	0.686679	0.610000	0.646072
50:50	0.764836	CD	0.882392	0.652659	0.750335
		HYP	0.702290	0.807018	0.751020
		MI	0.250696	0.543478	0.386266
		NORM	0.623188	0.754386	0.682540
		STTC	0.761364	0.783626	0.772334
60:40	0.775315	CD	0.246835	0.722222	0.367925
		HYP	0.705882	0.664032	0.684318
		MI	0.203187	0.714953	0.316443
		NORM	0.644024	0.486673	0.554400
		STTC	0.862545	0.732497	0.792219
70:30	0.766311	CD	0.203187	0.714953	0.316443
		HYP	0.623188	0.754386	0.682540
		MI	0.761364	0.783626	0.772334
		NORM	0.702290	0.807018	0.751020
		STTC	0.250696	0.543478	0.386266
80:20	0.791936	CD	0.743243	0.774531	0.627854
		HYP	0.894335	0.741667	0.830133
		MI	0.203187	0.714953	0.316443
		NORM	0.644024	0.486673	0.554400
		STTC	0.862545	0.732497	0.792219
90:10	0.774154	CD	0.668072	0.742389	0.703272
		HYP	0.300325	0.690299	0.418552
		MI	0.698182	0.45497	0.550933
		NORM	0.831119	0.855315	0.843043
		STTC	0.686679	0.610000	0.646072

Comparatively, the deep CNN architectures like InceptionTime and ResNet18 perform better compared to the light CNN model MobileNetV2, especially concerning the accuracy and F1 score for classifying the complex ECG signals. Regarding all types of models, it is observed that the NORM class is easiest to identify, while the class that is toughest to identify is always Hypertensive or HYP. This points towards the fact that more samples or some kind of data augmentation strategies are needed for the detection of the latter class. It can be observed that there is a point of saturation for the light models like MobileNetV2 after a certain point of increase in the proportion of the training samples.

The self-supervised learning models, MoCo v2 (ResNet34), BYOL (ResNet34), MAE (ResNet1D), and TS-JEPA (ResNet1D), were tested on the ECG database using the extracted features, which were classified using logistic regression, SVM, and random forest classifiers. Amongst the self-supervised learning models, the highest accuracy of 0.762 was obtained using logistic regression with MAE (ResNet1D), which performed significantly better than other self-supervised models. The accuracy of TS-JEPA was 0.654, whereas MoCo v2 and BYOL performed averagely with an accuracy of 0.576-0.614 using different classifiers. The results show that the performance of logistic regression is better than SVM and random forest classifiers, which implies that the features are mostly separable in the linear domain for ECG-related tasks.

TABLE IV
SELF-SUPERVISED LEARNING CLASSIFIER PERFORMANCE ON ECG DATASET

Model	Classifier	Accuracy
MoCo v2 (ResNet34)	Logistic Regression	0.614
MoCo v2 (ResNet34)	SVM	0.588
MoCo v2 (ResNet34)	Random Forest	0.601
BYOL (ResNet34)	Logistic Regression	0.608
BYOL (ResNet34)	SVM	0.576
BYOL (ResNet34)	Random Forest	0.567
MAE (ResNet1D)	Logistic Regression	0.762
MAE (ResNet1D)	SVM	0.750
MAE (ResNet1D)	Random Forest	0.729
TS-JEPA (ResNet1D)	Logistic Regression	0.654
TS-JEPA (ResNet1D)	SVM	0.651

The models with the highest accuracy in supervised learning, as well as self-supervised learning, in ECG classification are ResNet18 and MAE (ResNet1D), with an accuracy of 0.792 and 0.762, respectively. The performance of Resnet18 was impressive in all categories with a higher F1-score in both NORM and STTC, while HYP continued to be a difficult case. The use of self-supervised learning in the MAE with masked auto-encoding of a 1D ResNet model enabled the generation of a useful feature space which could be linearly separable with logistic regression reaching a competitive accuracy. While both models are capable of accurately detecting the patterns of both normal and normal twin conditions, conditions with subtle anomalies such as HYP continue to prove challenging.

CONCLUSION

In this work, supervised and self-supervised models based on deep learning for multi-class ECG classification are analyzed. Among the supervised models, ResNet18 had the best accuracy of 0.792 in classifying ECGs effectively by identifying intricate temporal features very accurately in classes such as NORM and STTC with minimal accuracy in subtle classes such as HYP. Among the self-supervised models, MAE (ResNet1D) had accuracy of 0.762 in distinguishing meaningful ECG features from unlabeled datasets. Thus, supervised models provide better raw results in ECG classification, whereas self-supervised models foster scalable learning techniques. Hence, leveraging both techniques might further enhance ECG classification results in the medical field.

REFERENCES

[1] Park, Junsang, Junho An, Jinkook Kim, Sunghoon Jung, Yeongjoon Gil, Yoojin Jang, Kwanglo Lee, and Il-young Oh. "Study on the use of standard 12-lead ECG data for rhythm-type ECG classification problems." Computer Methods and Programs in Biomedicine 214 (2022): 106521

[2] Soltanieh S, Hashemi J, Etemad A. In-distribution and out-of-distribution self-supervised ecg representation learning for arrhythmia detection. IEEE Journal of Biomedical and Health Informatics. 2023 Nov 10;28(2):789-800.

[3] Atienza, A., Bardram, J. and Puthusserypady, S., 2023. Subject-based Non-contrastive Self-Supervised Learning for ECG Signal Processing. arXiv preprint arXiv:2305.10347.

[4] Rahman, Atta-Ur, Rizwana Naz Asif, Kiran Sultan, Suleiman Ali Alsaif, Sagheer Abbas, Muhammad Adnan Khan, and Amir Mosavi. "ECG classification for detecting ECG arrhythmia empowered with deep learning approaches." Computational intelligence and neuroscience 2022, no. 1 (2022): 6852845.

[5] Gitau, Antony M., Ayoade Adeyemi, Belal Tavashi, and Bjørn-Jostein Singstad. "[Re] Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL." medRxiv (2025): 2025-01.

[6] Weimann, K. and Conrad, T.O., 2024. Self-Supervised Pre-Training with Joint-Embedding Predictive Architecture Boosts ECG Classification Performance. arXiv preprint arXiv:2410.13867.

[7] Weimann, K. and Conrad, T.O., 2024. Self-Supervised Pre-Training with Joint-Embedding Predictive Architecture Boosts ECG Classification Performance. arXiv preprint arXiv:2410.13867.

[8] Cheng, J., Zou, Q. and Zhao, Y., 2021. ECG signal classification based on deep CNN and BiLSTM. BMC medical informatics and decision making, 21(1), p.365.

[9] Weimann, K. and Conrad, T.O., 2021. Transfer learning for ECG classification. Scientific reports, 11(1), p.5251.