SUMMER OF MACHINE LEARNING AT SKOLTECH

# Improving Self Supervised Learning of ECG Signals Processing Via Encoder Embeddings Representation Enhancements


**Team:** Offline 18
**Offline participants:** Eva Bakaeva, Egor Padin
**Curator:** Konstantin Egorov

**Date of project start:** 17.07.2025

Skoltech Applied AI Center

# Project area and significance

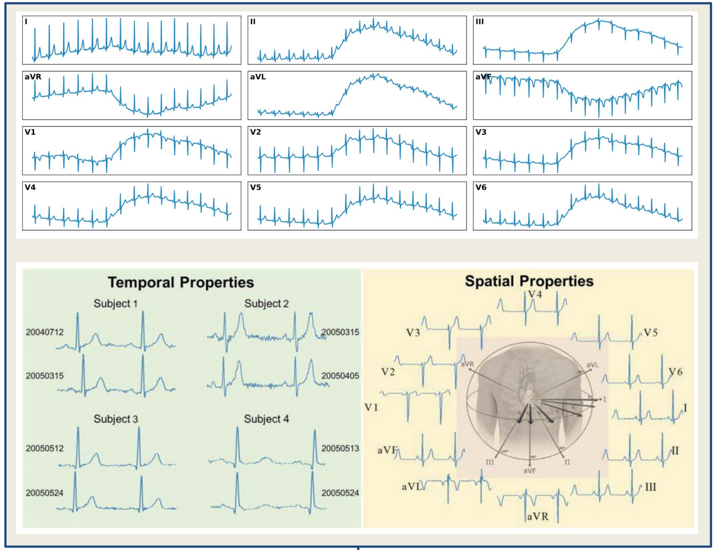
**Main Area Goal:** Prediction of the pathologies from Electrocardiogram (ECG)

**Significance of the Area:** Potential improvement for the medical assistance

**Area Problem**: Medical data is scarce and doesn't always include labeling

**Solution**: Self Supervised Learning (SSL) became an established method for biosignals processing

**Our Goal:** improve SSL pretrained models performance on downstream tasks




Model

↓

Diagnosis

1. top image from: https://doi.org/10.48550/arXiv.2410.08559
2. bottom image from: https://doi.org/10.1038/s41598-025-90084-2

# Problem statement and Hypothesis

| Problem | Hypothesis |
|---|---|
| Pure data-driven ECG embeddings frequently:<br>• Lack interpretability for clinical decision-making<br>• Fail to incorporate known electrophysiological biomarkers<br>• Demonstrate poor out-of-distribution performance on atypical cases | Enriching ECG embeddings with clinically validated ECG characteristics would enhance their representational quality and diagnostic utility |
| SSL pretrained models performance on classifications tasks might lack due to insufficient embedding representations learnt in embedding space | Improved latent space representations might lead to improved diagnoses classification |
| Standard ECG encoders produce:<br>• Continuous latent spaces where disease subtypes overlap<br>• Poor cluster separation for phenotypically similar but etiologically distinct conditions<br>• No natural mechanism for discrete disease categorization | Utilizing a VQ-VAE encoder could improve clustering of ECG embeddings from patients with shared symptoms, thereby enhancing latent space organization |

# Previous solutions (Baselines)

**Current SSL models:**

- **Joint-Emdedding Predicting Architecture (JEPA)**
- **Spatio-Temporal Masked Electrocardiogram Modeling (ST-MEM)**
- **A simple framework for contrastive learning of visual representations (SimCLR)**
- **Contrastive Multi-Segment Coding (CMCS)**
- **Contrastive Predictive Coding (CPC)**

**BASELINES**

# Datasets

| Name | Data size | Description | Labeled | Where to use |
|---|---|---|---|---|
| Shaoxing (Ningbo + Chapman) | 45152 records (45152 patients) | 10-second 12-lead ECG records from, recorded at 500Hz | 64 diagnostic labels | Pretrain |
| PTB-XL | 21837 records (18885 patients) | | 71 diagnostic labels, which are aggregated into five superclasses | Downstream tasks |

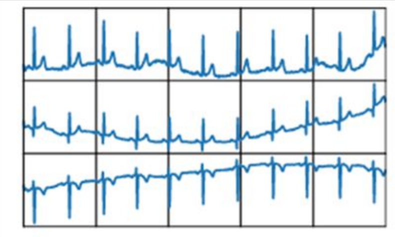
# Hybrid Input Representation

**Main idea:** We propose a dual-input SSL framework that jointly processes raw multi-lead ECG signals and precomputed clinical parameters.

For this purpose **ECG-JEPA** was used, that learns semantic representations of ECG data by predicting in the hidden latent space, bypassing the need to reconstruct raw signals. This approach offers several advantages in the ECG domain

Patches

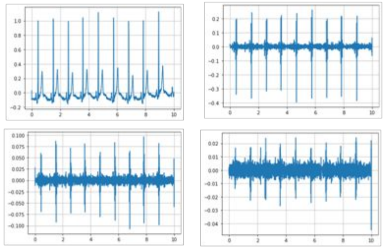
1. **ECG (8-lead)**



2. **Physiological features**

R-peaks, HR, Dynamic HR, STD, QRS, HRV STD, HRV RMS

3. **Wavelet transformation**



**Encoder (student)**

Patches representation

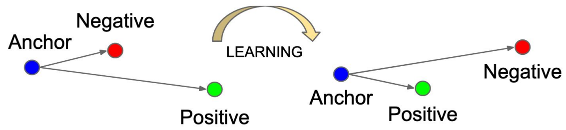
- ECG
- Physiological features
- Wavelet transformation

ECG representation

**Hypothesis**: Enriching ECG embeddings with clinically validated ECG characteristics would enhance their representational quality and diagnostic utility.

**Outcome:** Approach improved classification

# ST-MEM modification: Triplet-loss Finetuning




**Positive pairs:** same diagnosis
**Negative pairs:** different diagnoses

**Labels:** 5 General Diagnoses:
'CD', 'HYP', 'MI', 'NORM','STTC'

**Step 1:** Initialise TripletModel (ST_MEM_VIT freezed weights )

**Step 2:** Train unfreezed weights on training part of the PTB-XL dataset

**Step 3:** Use TripletModel for downstream tasks

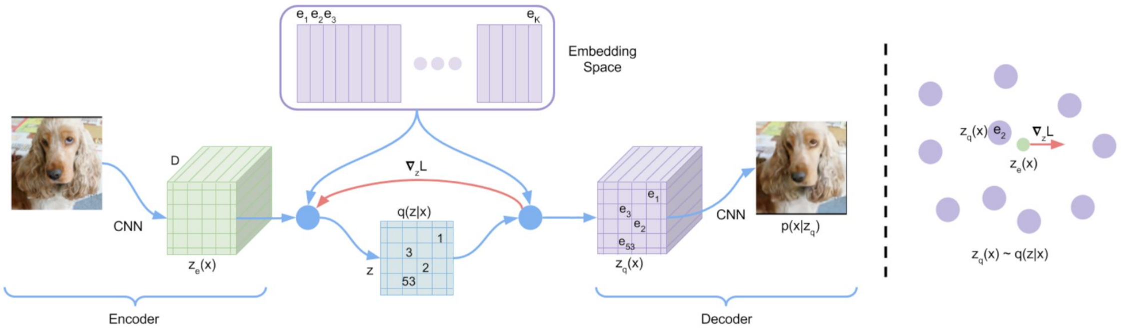
**Hypothesis:** Triplet loss would group similar inputs and distance different
**Outcome:** Method improved clustering and classification

```
TripletModel(
  (encoder): ST_MEM_ViT(
      seq_len=2250,
      patch_size=75,
      num_leads=12,
      num_classes=5,
      width=768,
      depth=12,
      mlp_dim=3072,
      heads=12,
      dim_head=64,
      qkv_bias=True,
      drop_out_rate=0.0,
      attn_drop_out_rate=0.0,
      drop_path_rate=0.0,
  )
  (fc1): Linear(in_features=768,
out_features=64, bias=True)
  (fc2): Linear(in_features=64,
out_features=9, bias=True)
)
```

1. image from: https://doi.org/10.48550/arXiv.1711.00937

# ST-MEM modification: VQ-VAE encoder






**Step 1:** Initialise VQ-VAE: encoder (ST_MEM_ViT freezed weights)

**Step 2:** Train unfreezed weights on sequence reconstruction task

**Step 3:** Use **ENCODER** of the model for downstream tasks

```
EncoderVQVAE(
  (encoder): ST_MEM_ViT(
    seq_len=2250,
    patch_size=75,
    num_leads=12,
    num_classes=None,
    width=768,
    depth=12,
    mlp_dim=3072,
    heads=12,
    dim_head=64,
    qkv_bias=True,
    drop_out_rate=0.0,
    attn_drop_out_rate=0.0,
    drop_path_rate=0.0,
  )
  (to_latent): Sequential(
    (0): Linear(in_features=768, out_features=512, bias=True)
    (1): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
    (2): Tanh()
  )
  (vq): VectorQuantizer()
```

**ENCODER**

```
  (decoder): Sequential(
    (0): Linear(in_features=512, out_features=768, bias=True)
    (1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
    (2): ReLU()
    (3): Dropout(p=0.1, inplace=False)
    (4): Linear(in_features=768, out_features=1536, bias=True)
    (5): ReLU()
    (6): Linear(in_features=1536, out_features=27000, bias=True)
  )
)
```

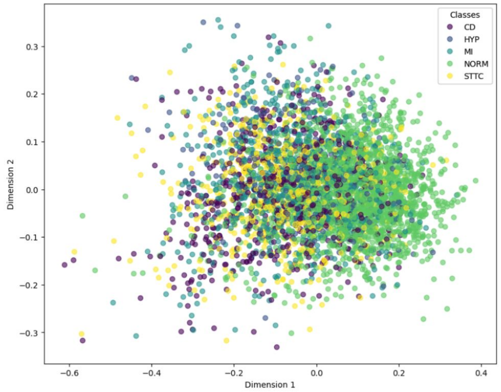
**Hypothesis:** Quantised latent space would improve clustering
**Outcome:** Method failed due to *codebook collapse*
**Potential solution:** Train whole encoder from scratch

1. image from: https://doi.org/10.48550/arXiv.1503.03832

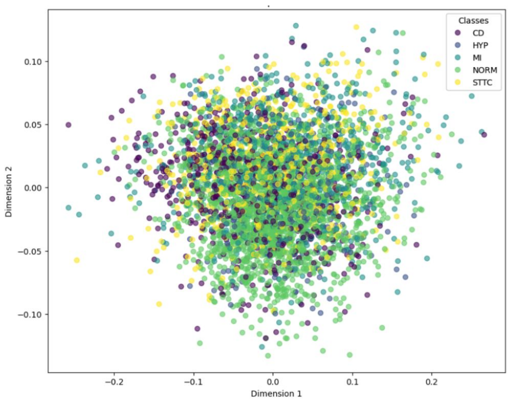
# Results: Encoders Clustering Performance

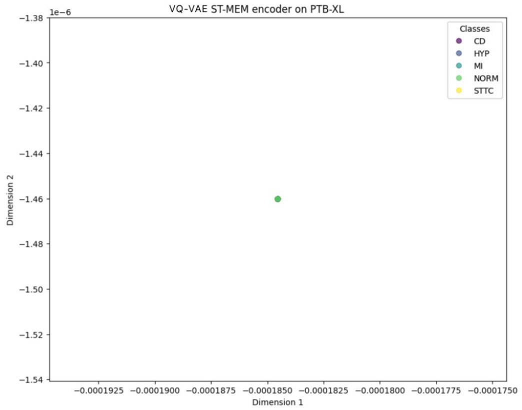
SUMMER OF MACHINE LEARNING AT SKOLTECH

ST-MEM Encoder on PTB-XL




ST-MEM Triplet-loss tuned Encoder on PTB-XL




| Method | DBI ↓ | Silhouette Score↑ | CHI↑ |
|---|---|---|---|
| *ST-MEM* | | | |
| Baseline | 9.690 | **-0.004** | **134.534** |
| VQ-VAE | *collapsed* | *collapsed* | *collapsed* |
| Triplet-loss | **6.755** | -0.054 | 73.565 |

**DBI** ~ How disorganized each cluster & how far clusters from each other
**Silhouette Score** ~ How well each point fits its cluster vs other clusters
**CHI** ~ Between-cluster variance opposed to  within-cluster variance vs

VQ-VAE: **codebook collapse** leads to mapping to a single embedding

Triplet-loss tuning **improved DBI**, but failed in Silhouette Score CHI

Both representations **don't perform well** in terms of clustering

# Results: Classification

SUMMER OF MACHINE LEARNING AT SKOLTECH

| Method | AUROC↑ | F1 Score↑ |
|---|---|---|
| *ECG-JEPA* | | |
| Baseline | 0.679 | 0.234 |
| Hybrid Input Representation | **0.784** | **0.462** |
| *ST-MEM linear probing* | | |
| Baseline | 0.695 | 0.138 |
| Triplet-loss | 0.686 | 0.135 |
| *ST-MEM KNN* | | |
| Baseline | 0.708 | 0.449 |
| VQ-VAE | *collapsed* | *collapsed* |
| Triplet-loss | 0.712 | 0.451 |

Main takeaways:

1. ECG-JEPA with Hybrid Input Representation is the **best performing model**

2. **Triplet-loss improves** ST-MEM performance compared to baseline when KNN classification used

3. None of the tested VQ-VAE training modifications overcame codebook collapse, therefore **full training is the only solution**

# Conclusions & Future Work

SUMMER OF MACHINE LEARNING AT SKOLTECH

- JEPA-ECG model with hybrid input representation **achieved superior performance**, demonstrating the effectiveness of combining joint-embedding architectures with **ECG-specific adaptations**

- Triplet Loss fine-tuning **enhanced DBI clustering score** of the embeddings and **classification performance** on the diagnosis prediction task

- Overall **clustering performance still lacks** among the examined encoders and to be improved

- Further experiments are necessary to determine whether **additional computational resources** used would improve results further or not

- Evaluate the performance of the **fully trained** proposed methods rather than the fine-tuned versions

- Future work should expand testing to **additional datasets and baselines** to validate the robustness of our modifications

- A **dedicated hyperparameter search** for optimal configurations is warranted in subsequent studies

# Team

SUMMER OF MACHINE LEARNING AT SKOLTECH

**Eva Bakaeva**
MIPT Bachelors Graduate,
Cognitive Modelling Center Staff

**Project responsibilities:**

1. Triplet-loss Finetuning ST-MEM modification
2. VQ-VAE encoder ST-MEM modification
3. Embedding clustering analysis

@HARUSP9X

**Egor Padin**
Tyumen Petroleum Research Center

**Project responsibilities:**

1. Hybrid Input Representation (Clinical Features) for Encoder JEPA modification

@EGOR_PADIN