Diversity of Synthetic Pretraining Objectives and Robustness in Tabular Foundation Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does varying the diversity of synthetic pretraining objectives impact the robustness of tabular foundation models on out-of-distribution subsets of TabBench. 13 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Pretraining Strategies and Scaling for ECG Foundation Models: A Systematic Study. Research question: How does varying the diversity of synthetic pretraining objectives impact the robustness of tabular foundation models on out-of-distribution subsets of TabBench?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.5/10.

3 Results

15 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 5.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The study covers five different pretraining methodologies trained on a corpus comprising over 11 million ECG samples.	×	0.09
State space models (SSM) are confirmed as the superior architecture choice across all pretraining paradigms compared to	×	0.10
Among the five pretraining methodologies evaluated, CPC shows the strongest and most transferable representations across	✓	0.15
The data2vec pretraining methodology consistently lags behind other methodologies across all evaluation modes and scalin	×	0.06
Scaling behavior is most clearly identified for the CPC and JEPA pretraining methodologies.	×	0.13
Lower pretraining loss correlates with small residual errors in downstream tasks.	×	0.03
Structured state space models have shown superior performance on long sequences in supervised ECG settings in prior stud	×	0.13
All models in the study operate at a sampling rate of 240 Hz on 12-lead ECG inputs.	×	0.03
The S4 backbone with a model dimension of 512 consistently outperforms larger dimensions (768, 1024) and alternative bac	×	0.01
The study adopts a 4-layer S4 backbone with dimension 512 as the default architecture based on ablation results.	×	0.03
The CNN stem used in the encoder consists of four convolutional layers with batch normalization.	×	0.01
The Transformer backbone variant utilizes RoPE positional encoding and GELU activations.	×	0.00
The data2vec objective trains the model to predict the EMA teacher's averaged top-k contextualized layer representations	×	0.02

References

- http://arxiv.org/abs/2601.04110v2

- http://arxiv.org/abs/2402.01204v4
- http://arxiv.org/abs/2605.12241v1