Tabular Foundation Models vs. Gradient Boosting in Few-Shot Learning Accuracy

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: To what extent do tabular foundation models maintain prediction accuracy compared to gradient boosting methods when evaluated on few-shot learning benchmarks with limited labeled rows. This study compared the performance of classical feature-based machine learning models (CMLs) and large language models (LLMs) in predicting COVID-19 mortality using high-dimensional tabular data from 9,134 patients across four hospitals. Seven CML models, including XGBoost and. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Large Language Models versus Classical Machine Learning: Performance in COVID-19 Mortality Prediction Using High-Dimensional Tabular Data. Research question: To what extent do tabular foundation models maintain prediction accuracy compared to gradient boosting methods when evaluated on few-shot learning benchmarks with limited labeled rows?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

14 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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 dataset consists of 9,057 total patients, with 1,818 mortality cases and 7,239 survivors.	×	0.03
The mean age of the total patient cohort (N=9,057) is 58.4 years with a standard deviation of 19.8.	×	0.02
In the study's cross-validation, the Classical Machine Learning (CML) Random Forest (RF) model achieved an accuracy of 0	×	0.11
In the study's cross-validation, the Classical Machine Learning (CML) Random Forest (RF) model achieved an AUC of 0.70 $\pm$	×	0.11
In the study's cross-validation, the Classical Machine Learning (CML) Support Vector Machine (SVM) model achieved an acc	×	0.07
In the study's cross-validation, the Classical Machine Learning (CML) Logistic Regression (LR) model achieved an accurac	×	0.08
In the study's cross-validation, the Classical Machine Learning (CML) Decision Tree (DT) model achieved a specificity of	×	0.06
In the Hegselmann (2023) TabLLM study on the Diabetes dataset, the zero-shot AUC performance was 0.82.	×	0.05
In the Hegselmann (2023) TabLLM study on the Heart dataset, the zero-shot AUC performance was 0.54.	×	0.05
In the Wang (2024) MediTab study using the MIMIC dataset, the Random Forest (RF) model achieved an accuracy of 0.78 with	×	0.05
In the CRADLE benchmark, GPT-4 achieved a zero-shot accuracy of 0.21 and an F1 score of 0.22.	×	0.10
In the CRADLE benchmark, GPT-3.5 achieved a zero-shot accuracy of 0.56 and an F1 score of 0.52.	×	0.10
In the CRADLE benchmark, the Random Forest (RF) model achieved an accuracy of 0.80 and an F1 score of 0.57 with a traini	×	0.07

References

- http://arxiv.org/abs/2409.02136v2
- http://arxiv.org/abs/1910.03560v2
- http://arxiv.org/abs/2506.13817v1