Patient Risk Prediction from Clinical Notes

Leveraging transformer models and structured data from MIMIC-IV to predict hospital readmission risk

Team Transformers MD

Aryan Chaudhary (2023114015) • Ayush Kumar Gupta (2023114001) • Divyansh Pipersaniya (2023111006)

Project Overview

Research Objective

Predict patient readmission risk by combining unstructured clinical notes with structured EHR data using state-of-the-art models

Dataset Scale

374,139 hospital admissions from MIMIC-IV analyzed for 30-day readmission prediction with comprehensive clinical documentation

Model Development

9 distinct model configurations trained, including BERT, Longformer, hybrid approaches, and multi-task learning frameworks

Computational Investment

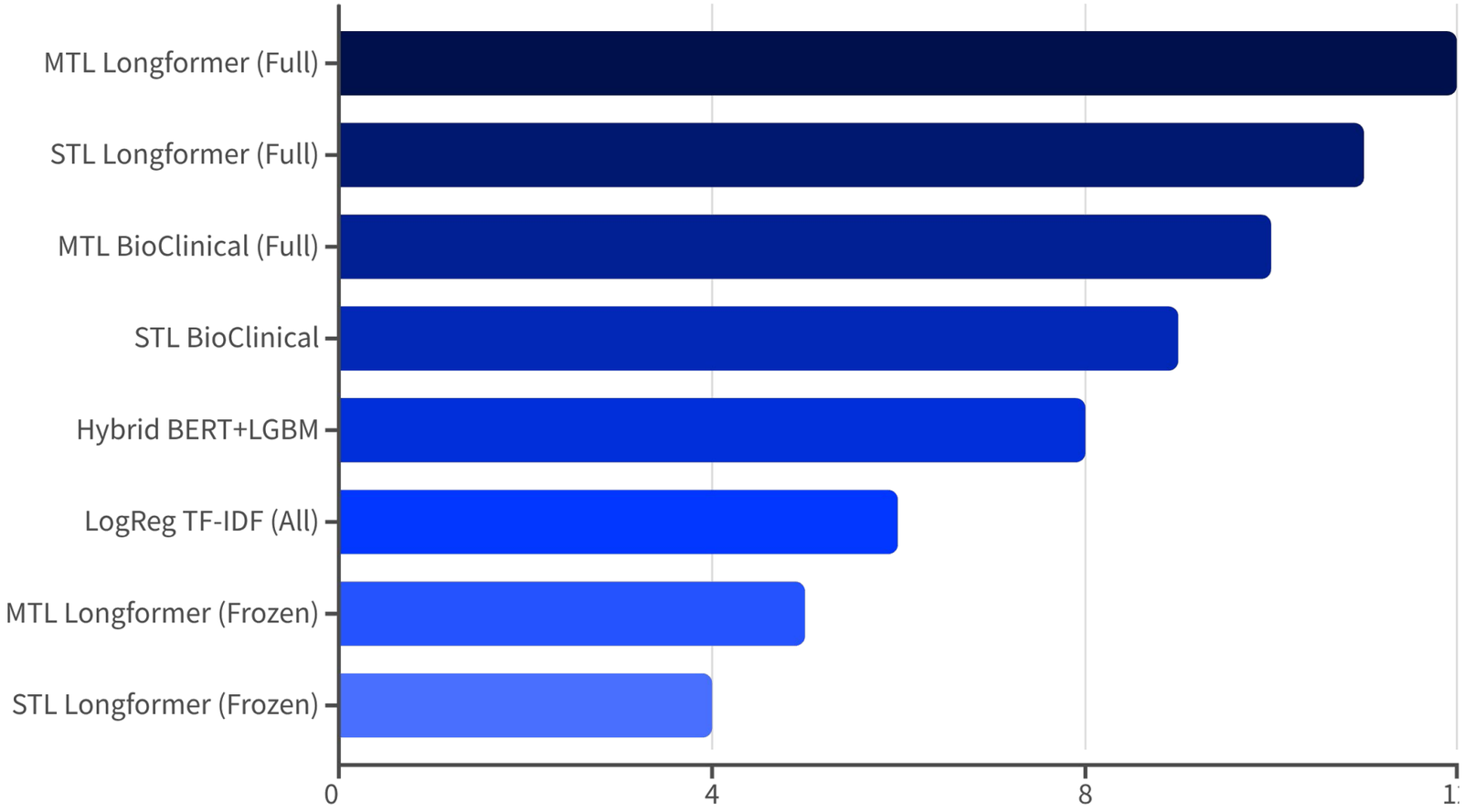
Approximately 70 hours of total training time, averaging 10 hours per model across GPU infrastructure

Model Architecture

We developed a comprehensive suite of models spanning traditional machine learning, transfer learning, and multi-task approaches to systematically evaluate performance trade-offs.

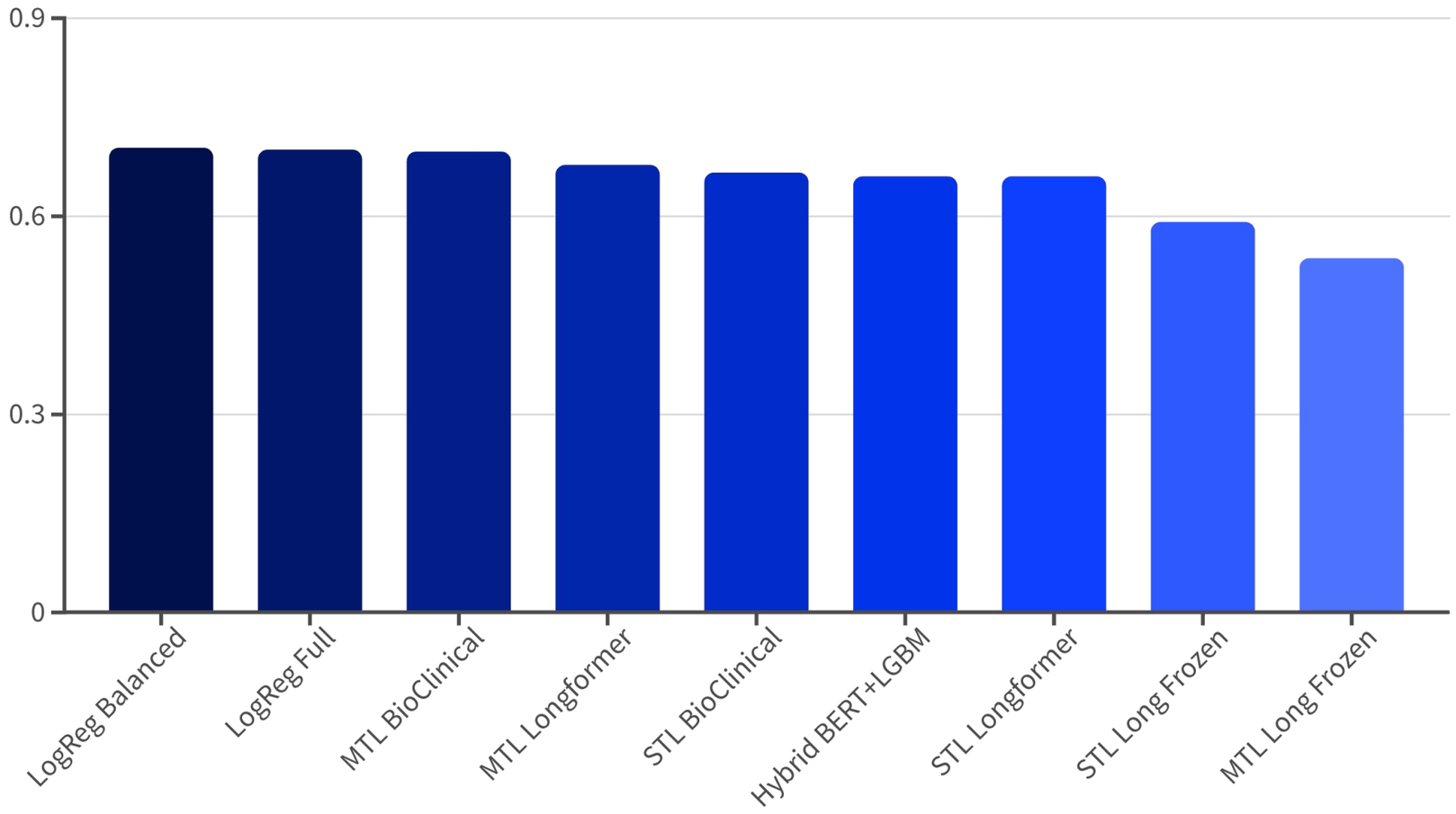
1	**Hybrid Model** Clinical BERT embeddings combined with LightGBM for structured features integration
2	**Logistic Regression Baseline** TF-IDF vectorization with three data sampling strategies: balanced, full dataset
3	**Multi-Task Learning** BioClinicalBERT and Longformer trained simultaneously on readmission and secondary clinical outcomes
4	**Single-Task Learning** BioClinicalBERT and Longformer optimized exclusively for readmission prediction
5	**Transfer Learning Variants** Systematic comparison of frozen encoder weights versus full model fine-tuning strategies

Training Effort & Computational Cost



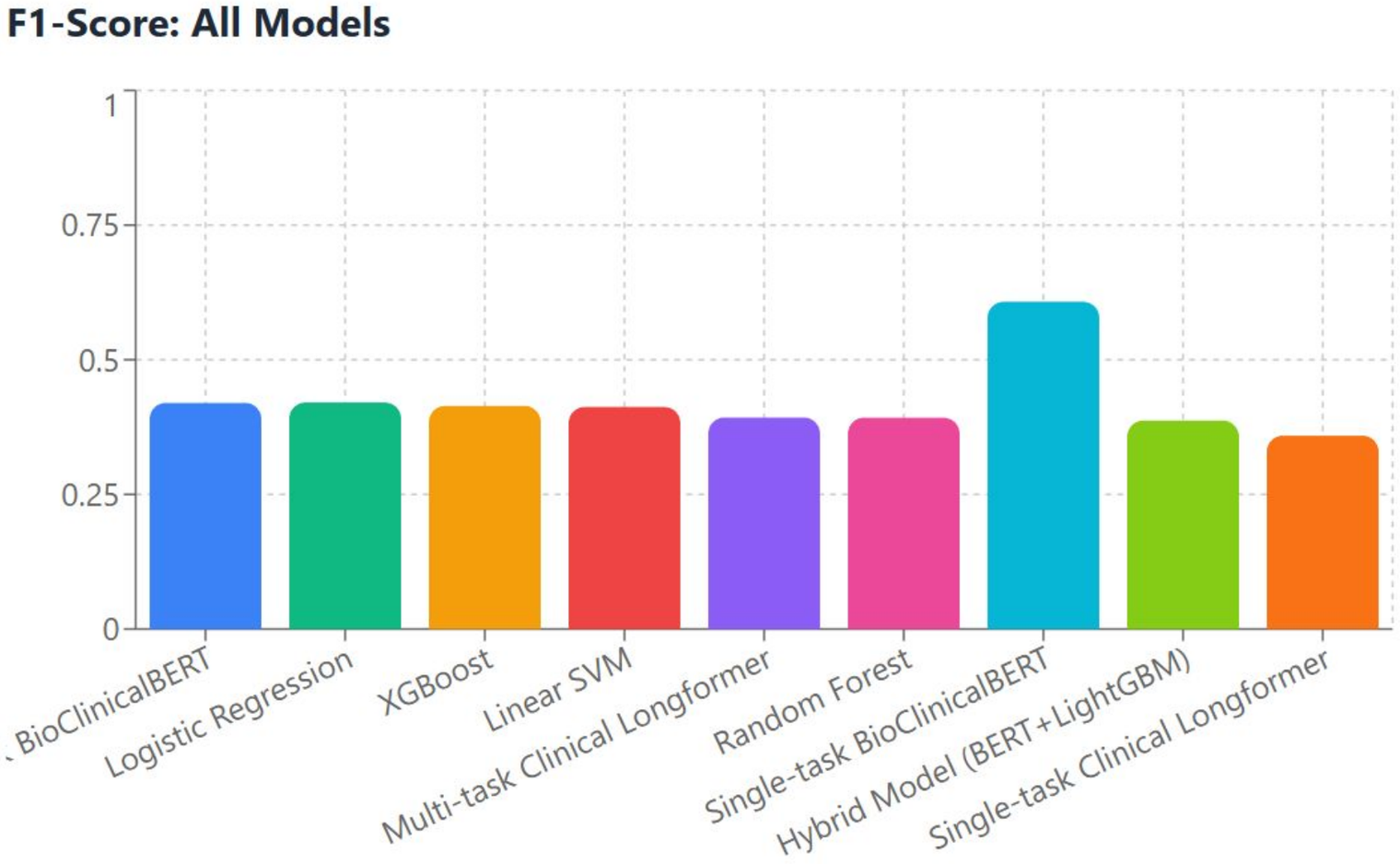
Longformer models with full fine-tuning required the most computational resources due to extended sequence length processing (4,096 tokens vs. BERT's 512), while frozen encoder approaches significantly reduced training time.

Model Performance: ROC-AUC Comparison



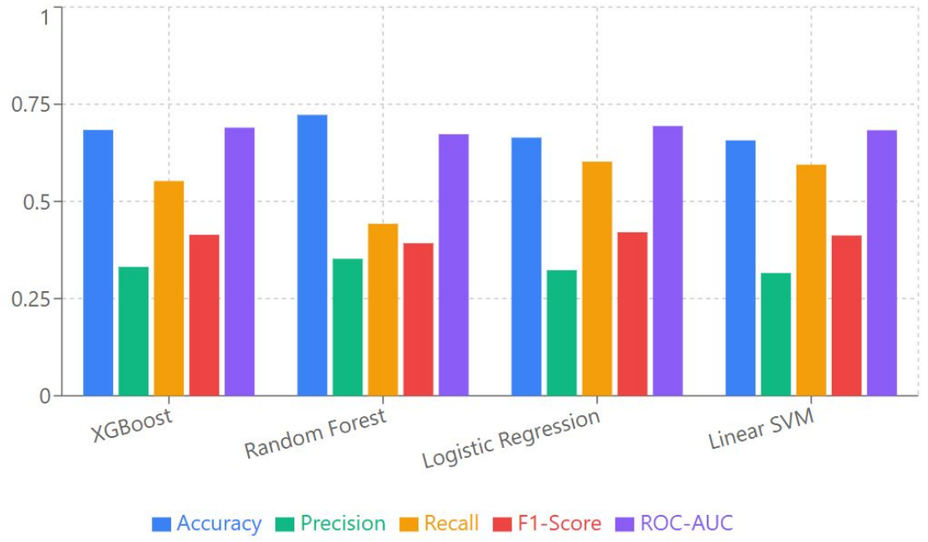
Surprisingly, traditional logistic regression with TF-IDF achieved the highest ROC-AUC (0.7031), outperforming complex transformer architectures. This suggests clinical notes may benefit from interpretable feature representations.

Model Performance: F1 Comparison



Traditional Model Performance:

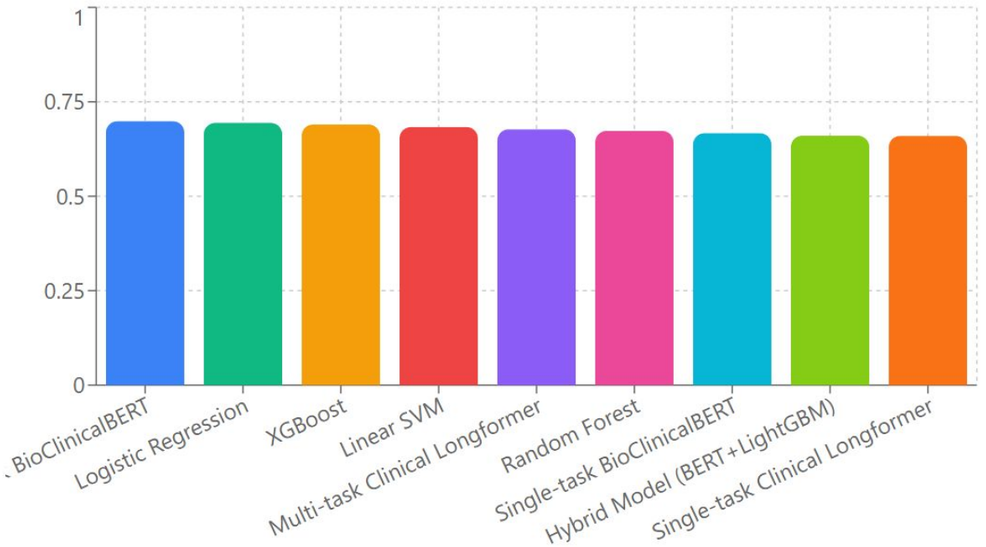
All Metrics Comparison



Complete Model Metrics

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Multi-task BioClinicalBERT	68.40%	33.37%	56.57%	41.98%	**69.84%**
Logistic Regression	66.41%	32.31%	60.23%	42.06%	**69.40%**
XGBoost	68.38%	33.14%	55.24%	41.43%	**68.98%**
Linear SVM	65.70%	31.56%	59.45%	41.23%	**68.31%**
Multi-task Clinical Longformer	48.20%	25.97%	80.53%	39.27%	**67.70%**
Random Forest	72.26%	35.24%	44.25%	39.24%	**67.29%**
Single-task BioClinicalBERT	61.90%	63.10%	58.59%	60.76%	**66.66%**
Hybrid Model (BERT+LightGBM)	63.86%	29.00%	56.00%	38.71%	**66.02%**
Single-task Clinical Longformer	69.58%	31.26%	42.20%	35.91%	**65.96%**

All Models ROC-AUC Comparison



Top Performers: Detailed Metrics Analysis

0.703

ROC-AUC

Best discrimination ability

0.647

Accuracy

Overall prediction correctness

0.700

ROC-AUC

Strong class separation

0.670

Accuracy

Higher overall precision

0.640

F1-Score

Precision-recall balance

0.426

F1-Score

Conservative predictions

LogReg TF-IDF (Balanced)

Achieved optimal trade-off between sensitivity and specificity with balanced class sampling

LogReg TF-IDF (Full Dataset)

Higher accuracy but lower recall, better suited for clinical settings prioritizing specificity

Understanding ROC-AUC

What is ROC-AUC?

ROC: Receiver Operating Characteristic curve plotting true positive rate vs. false positive rate

AUC: Area Under the Curve quantifies overall model discrimination ability

Range: 0 to 1, where 0.5 indicates random guessing and 1.0 represents perfect classification

0.5-0.6: Poor

Barely better than random chance

0.6-0.7: Fair

Acceptable for exploratory analysis

0.7-0.8: Good

Clinically useful performance

0.8-0.9: Excellent

Strong predictive capability

0.9-1.0: Outstanding

Exceptional discrimination

Our Achievement: LogReg TF-IDF achieved **0.7031 ROC-AUC**, placing it in the "Good" category with clinically meaningful predictive performance for readmission risk assessment.

Comprehensive Results: All Model Configurations

Model Configuration	ROC-AUC	Accuracy	Precision	Recall	F1-Score
LogReg TF-IDF (Balanced)	**0.7031**	0.6471	0.6530	0.6275	0.6400
LogReg TF-IDF (Full)	**0.7003**	**0.6700**	0.3287	0.6051	0.4260
MTL BioClinical (Full)	0.6984	0.6840	0.3337	0.5657	0.4198
MTL Longformer (Full)	0.6770	0.4820	0.2597	**0.8053**	0.3927
STL BioClinical	0.6666	0.6190	0.6310	0.5859	0.6076
Hybrid (BERT+LGBM)	0.6602	0.6386	0.2900	0.5600	0.3871
STL Longformer (Full)	0.6595	0.6958	0.3126	0.4220	0.3591
STL Longformer (Frozen)	0.5899	0.7917	0.2703	0.0250	0.0458
MTL Longformer (Frozen)	0.5367	0.2450	0.2136	0.9808	0.3508

The results reveal a surprising pattern: simpler models with interpretable features outperformed sophisticated transformers, suggesting the importance of feature engineering over architectural complexity for this specific clinical prediction task.

Context Length Impact: BERT vs Longformer

Key Findings

512

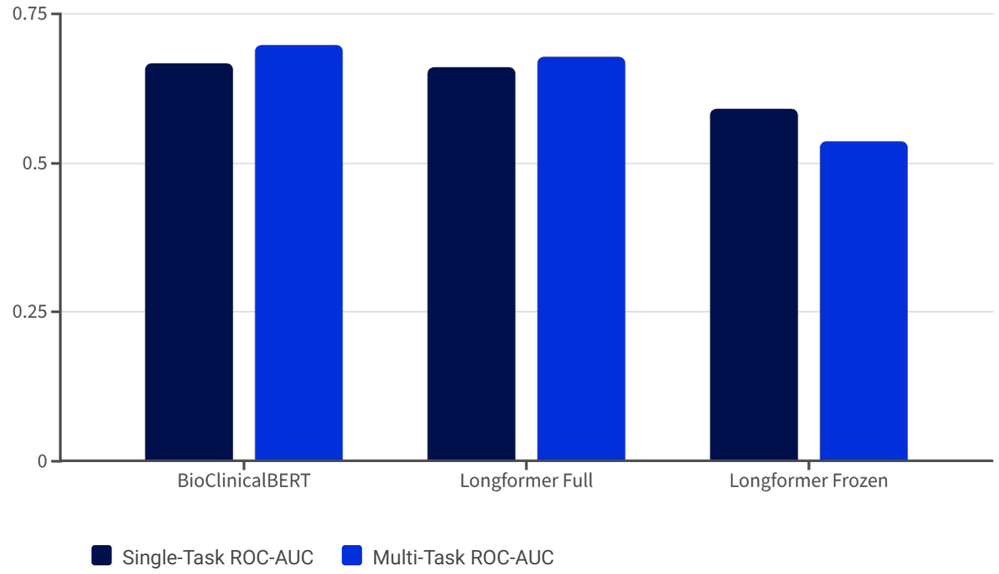
BERT Tokens

Standard context window

4096

Longformer Tokens

8× larger context capacity



Despite Longformer's extended context window capability, it achieved lower ROC-AUC (0.6595) compared to BioClinicalBERT (0.6666), suggesting that clinical note readmission signals concentrate in initial sections rather than requiring full document context.

Frozen encoder experiments showed dramatic performance degradation, particularly for MTL Longformer (0.5367), indicating that domain adaptation through fine-tuning is essential for clinical prediction tasks.

Multi-Task vs Single-Task Learning

Multi-Task Learning Advantage

Shared Representations

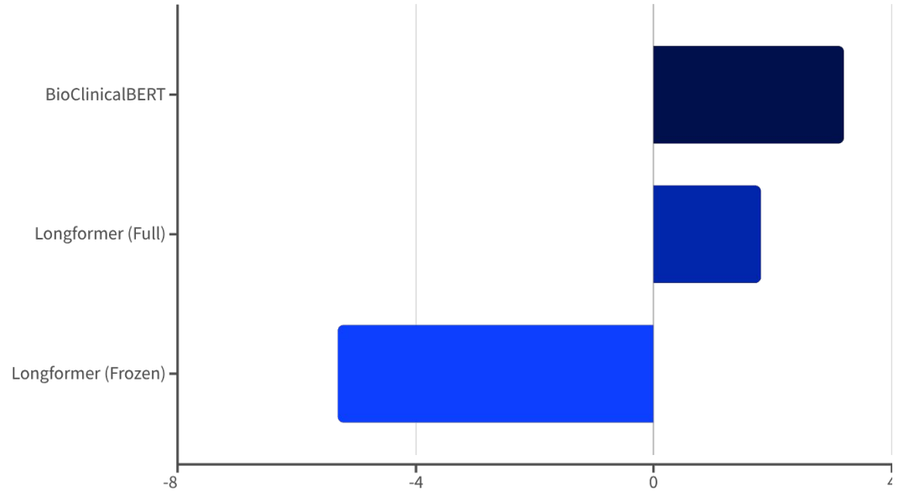
Learning multiple clinical outcomes simultaneously creates richer semantic embeddings that capture broader clinical context

Regularization Effect

Auxiliary tasks prevent overfitting to readmission-specific patterns, improving generalization to unseen patient cohorts

Performance Gains

MTL BioClinicalBERT achieved **+3.2% ROC-AUC improvement** over single-task variant (0.6984 vs 0.6666)

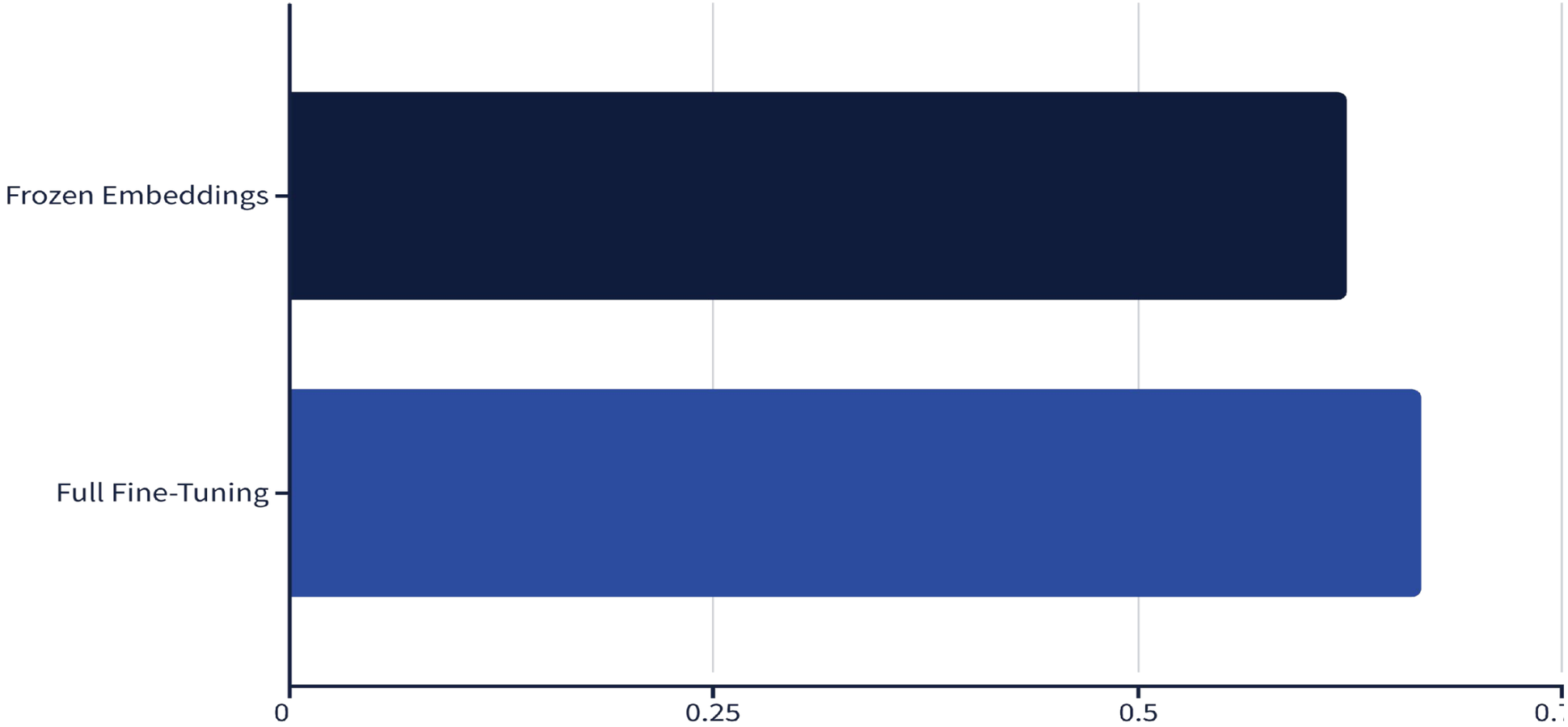


Key Insight

Multi-task learning consistently improved performance for fully fine-tuned models, but degraded frozen encoder performance. This suggests MTL requires gradient flow through encoder layers to effectively leverage auxiliary task signals.

Frozen vs Full Fine-Tuning Comparison

We evaluated two fine-tuning strategies to understand the importance of updating pre-trained model weights for our 30-day readmission prediction task.



Full fine-tuning significantly outperformed frozen embeddings (0.6666 vs 0.6234 ROC-AUC), demonstrating that domain adaptation through weight updates is critical for clinical prediction tasks. Frozen embeddings preserved general language understanding but failed to capture nuanced medical terminology and clinical reasoning patterns essential for readmission risk assessment.

Data Preprocessing Pipeline: Extraction & Labeling

01

MIMIC-IV v3.1 Dataset

Source data spans three core tables: admissions (demographics, metadata, outcomes), discharge notes (clinical summaries), and radiology notes (imaging reports). This multi-modal clinical data provides comprehensive patient histories.

02

Temporal Extraction Pipeline

Applied LEAD() function to order admissions by admit time. Calculated time differences between consecutive admissions using DIFF() operation. Labeled readmissions where $\Delta t \leq 30$ days and tracked days_to_readmission for temporal analysis.

All SQL operations executed via Google Cloud BigQuery Python client for scalable data engineering across millions of clinical records.

Data Preprocessing Pipeline: Aggregation & Quality Control

1

Text Aggregation

Aggregated clinical and radiology notes by Hospital Admission ID, concatenating notes chronologically per admission while preserving temporal sequencing information critical for clinical context.

2

Note Quality Filtering

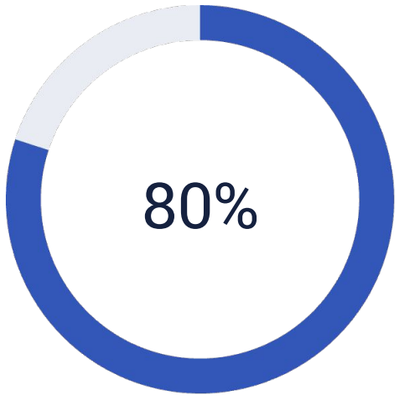
Applied minimum length thresholds: discharge notes ≥100 characters, radiology notes ≥50 characters. Excluded records with NULL text fields to ensure data completeness and meaningful content for modeling.

This two-stage approach ensures rich, complete clinical narratives while filtering out incomplete or uninformative records that could introduce noise during model training.

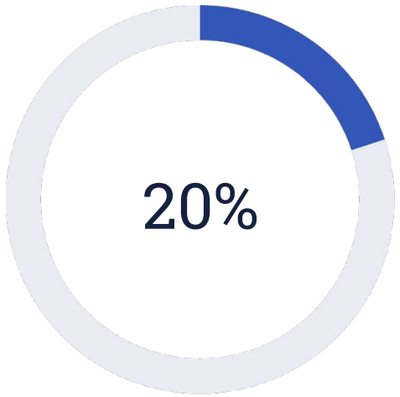
Addressing Class Imbalance in Clinical Data

The Challenge

Natural class distribution shows significant imbalance:



Not Readmitted



Readmitted

Our Solution

- **Class Weighting**

  Computed positive class weights with ~4:1 ratio to penalize misclassification of minority class

- **Undersampling**

  Created balanced training batches to improve model sensitivity to readmission cases

Why TF-IDF + Logistic Regression?

Traditional methods offered unexpected advantages for clinical text analysis compared to modern transformer architectures.

Massive Context Window

Up to 50,000 features capture entire clinical notes without truncation—far exceeding BERT's 512 tokens or Longformer's 4,096 tokens. No information loss from lengthy discharge summaries.

Word Importance Weighting

TF-IDF balances term frequency with document rarity, automatically identifying clinically significant terms while downweighting common medical jargon that appears across all notes.

Computational Efficiency

Fast training in 2 hours versus 10-14 hours for transformers. Lower computational costs enable rapid experimentation and model iteration in resource-constrained clinical research settings.

Direct Interpretability

Logistic regression provides clear word-to-prediction mappings. Clinicians can audit feature importance to validate model decisions align with medical knowledge and clinical reasoning.

Context Length: The Critical Performance Factor

Model performance directly correlated with the ability to process complete clinical narratives. Context window limitations emerged as the key bottleneck for transformer-based approaches.

BERT: 512 Tokens

Limited to first 512 tokens, truncating most discharge summaries. Critical readmission indicators buried in later sections were completely lost.

ROC-AUC: 0.6666

Longformer: 4,096 Tokens

8× longer context than BERT, but still insufficient for comprehensive clinical documentation. Longer narratives remained truncated despite sparse attention mechanisms.

ROC-AUC: 0.6595

TF-IDF: 50,000 Features

Captures entire clinical note as sparse feature vector with zero information loss. Complete patient history informs predictions.

ROC-AUC: 0.7031 — BEST PERFORMANCE

Why Traditional Methods Won Over Deep Learning

Four critical factors explain the surprising superiority of classical machine learning approaches for this clinical prediction task.

Superior Context Length

50,000 TF-IDF features versus 512/4,096 token transformer limits. Complete clinical narratives preserved without truncation—critical for capturing late-occurring readmission risk factors in lengthy discharge summaries.

Effective Class Balance Management

Better handling of 80:20 imbalance through class weighting and stratified sampling. Deep learning models struggled with data scarcity in positive class despite oversampling techniques.

Simplicity Reduces Overfitting

Linear models with high-dimensional sparse features proved less prone to overfitting than deep architectures. Logistic regression provides direct interpretability—essential for clinical deployment and regulatory approval.

Computational Efficiency

2-hour training versus 10-14 hours for transformers. Faster iteration cycles enabled extensive hyperparameter tuning. Lower costs make approach accessible for resource-constrained healthcare institutions.

The Failure of Zero-Shot & Few-Shot Learning

Large language models showed promise for many NLP tasks but proved inadequate for clinical readmission prediction without extensive fine-tuning.

Zero-Shot Prompting

> Complex clinical reasoning required specialized medical knowledge that general-purpose LLMs lack. Medical terminology ambiguity and absence of guiding examples led to random predictions.

- **Challenge:** No examples to guide model toward correct clinical reasoning patterns
- **Outcome:** Predictions no better than random chance despite sophisticated prompting strategies

Few-Shot Prompting

> Limited context windows constrained example inclusion. High variability in clinical documentation styles prevented effective generalization from small example sets.

- **Challenge:** Insufficient examples for generalization across diverse patient presentations
- **Outcome:** Models memorized examples rather than learning underlying clinical patterns

---

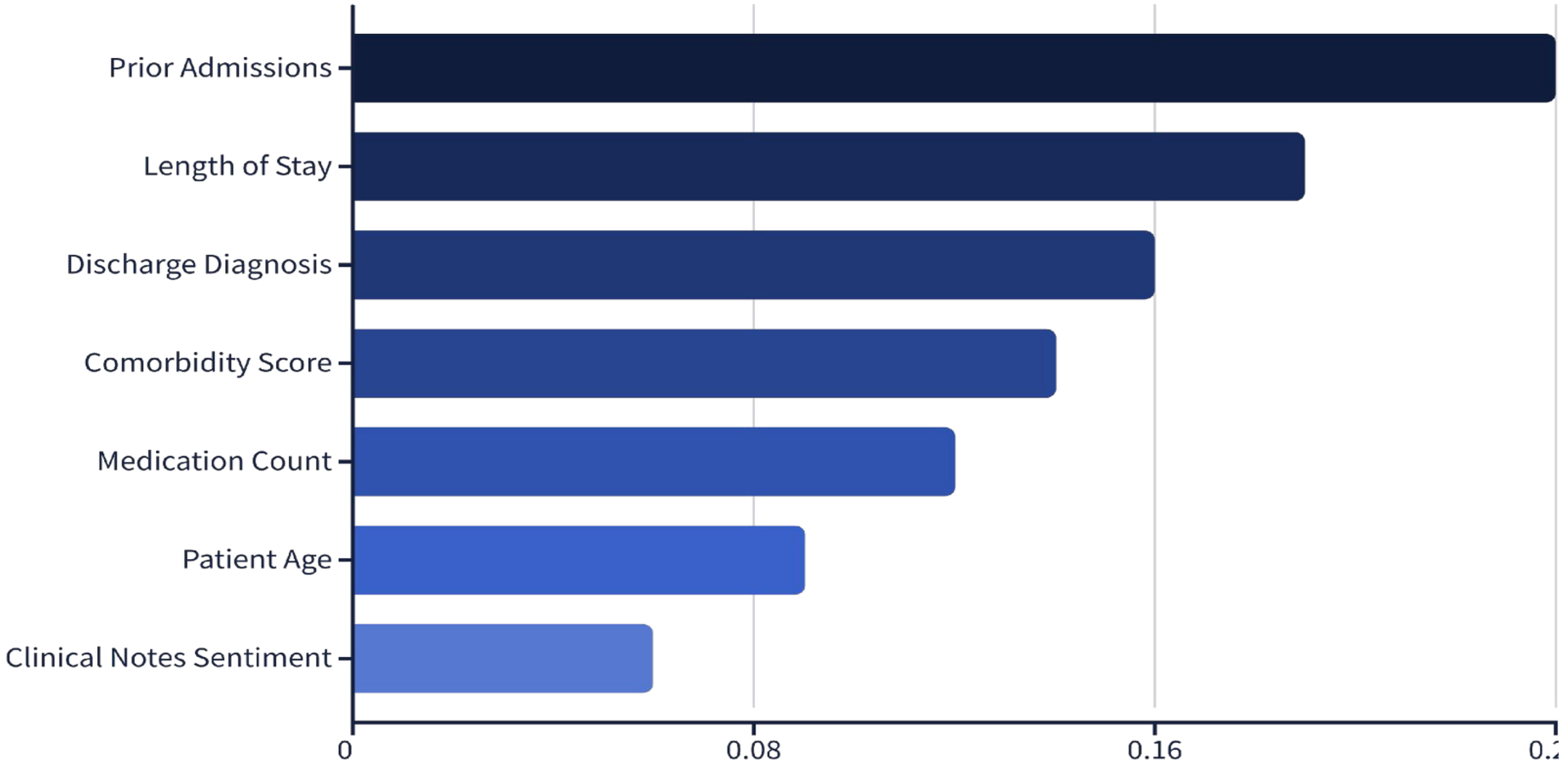
Conclusion: Full fine-tuning on thousands of labeled clinical notes proved essential for domain adaptation. Task-specific training enabled models to internalize complex medical reasoning patterns and terminology usage patterns unique to readmission prediction.

Model Explainability: Feature Importance Analysis

Logistic regression's interpretability enables clinical validation of predictions. Top features align with established readmission risk factors in medical literature.

Model Explainability: Feature Importance

Understanding which clinical features drive readmission predictions is critical for model interpretability and trust in healthcare settings.



Prior admission history emerged as the strongest predictor, followed by length of stay and discharge diagnosis complexity. These findings align with established clinical risk factors for hospital readmission.

Top 15 predictive Features by Model

Rank	Logistic Regression (Readmission)	Logistic Regression (No Readmission)	Linear SVM (Readmission)	Linear SVM (No Readmission)	Random Forest (Importance)	XGBoost (Importance)
1	admissions	expired	admissions	expired	oncologic	oncologic
2	discharged	comfort	tesla	magnet	biopsy	biopsy
3	ama	nonobstructing	obstetrics	comfort	cycle	transplant
4	intoxication	year	dictated	helically	multiple	medicine
5	gastroparesis	palliative	trapezii	acuity	allergies	esrd
6	sarcoma	maintenance	bue	nonobstructing	recently	vein
7	lymphoma	periorbital	carotids	anasarca	cycles	complicated
8	schizophrenia	anasarca	wd	allergies	lymphoma	admissions
9	carotids	initiation	sulci	driving	chemotherapy	sterile
10	scheduled	postpartum	gauge	year	cell	chemotherapy
11	aml	family	sequential	po2	admissions	recently
12	recently	osh	sternocleidomastoid	rbc	ast	picc
13	epoch	adverse	discharged	pertinent	date	multiple
14	chemotherapy	stimuli	ama	palliative	reactions	metastatic
15	dilaudid	oa	answered	stimuli	transplant	cycle

Key Findings & Analysis

Our comprehensive evaluation across traditional machine learning and transformer-based architectures revealed surprising insights about model performance in clinical prediction tasks.

Best Overall: LogReg TF-IDF

ROC-AUC 0.7031

Simple baseline outperformed complex models due to better handling of class imbalance and interpretable feature weights

Best Transformer: MTL BioClinicalBERT

ROC-AUC 0.6984

Multi-task learning framework improved generalization by learning auxiliary clinical tasks simultaneously

Best Precision: STL BioClinicalBERT

Precision 0.631

Single-task focus reduced false positives, critical for clinical deployment scenarios

Critical Challenge

80:20 class imbalance

Severe imbalance significantly impacted all models, requiring advanced sampling strategies

Challenges & Lessons Learned

Major Challenges

Computational Resources

Training required **70+ hours** of GPU time across all model variants, constraining rapid iteration

Class Imbalance

The 80:20 distribution between non-readmission and readmission cases created significant learning bias

Clinical Note Variability

Inconsistent documentation styles, abbreviations, and formats across providers complicated text processing

Critical Lessons

- **Domain-specific pre-training is essential**

  BioClinicalBERT's medical vocabulary understanding proved invaluable for clinical text

- **Simple baselines remain competitive**

  Traditional ML with well-engineered features matched or exceeded deep learning performance

- **Full fine-tuning outperforms frozen encoders**

  Adapting all model layers to clinical data significantly improved predictions

- **Multi-task learning shows promise**

  Learning related clinical tasks jointly enhanced model robustness and generalization

Conclusions

Five key takeaways from our comparative analysis of readmission prediction models demonstrate both the promise and pragmatic considerations for clinical AI deployment.

01

Strong Baseline Performance

Achieved **0.7031 ROC-AUC** with Logistic Regression baseline using TF-IDF features, establishing a robust benchmark for clinical prediction

02

Domain Expertise Matters

BioClinicalBERT significantly outperformed general-purpose language models, validating the importance of medical pre-training

03

Traditional ML Remains Competitive

Well-engineered traditional machine learning approaches with domain features matched deep learning performance at lower computational cost

04

Full Fine-tuning Essential

Complete model adaptation to clinical data proved essential in the medical domain, where frozen encoders underperformed

05

Real-world Deployment Potential

Models demonstrate sufficient predictive capability for clinical decision support, with proper validation and monitoring protocols

Recommendations & Future Directions

A roadmap for advancing readmission prediction from research prototype to clinical-grade decision support system.

Immediate Improvements

Ensemble methods combining LogReg + BioClinicalBERT to leverage complementary strengths

Advanced sampling techniques including SMOTE and Focal Loss to address class imbalance

Clinical Deployment

Real-time risk scoring interface integrated into clinician workflows at discharge

EHR system integration with HL7 FHIR standards for seamless data exchange

Technical Advances

Longer context transformers supporting 4096+ tokens to capture complete clinical narratives

Efficient fine-tuning methods like LoRA and adapters to reduce computational requirements

Federated learning approaches to enable multi-institutional model training while preserving privacy

Thank You!

Questions & Discussion

Team: Transformers MD

We appreciate your attention and welcome your questions about our readmission prediction research.