Patient Readmission Detection using Clinical NLP and Multi-Task Learning

Aryan Chaudhary **Ayush Kumar Gupta** **Divyansh Pipersaniya**
International Institute Of Information Technology, Hyderabad

Abstract

Reducing 30-day hospital readmissions remains a persistent challenge in clinical care, demanding accurate and scalable predictive models capable of leveraging the richness of Electronic Health Records (EHRs). In this work, we conduct a comprehensive, multi-stage evaluation of machine-learning and deep-learning approaches for early readmission prediction using the MIMIC-IV dataset, with a primary focus on the comparative modeling capabilities of classical algorithms, transformer-based architectures, and Multi-Task Learning (MTL) frameworks.

We develop a unified pipeline that processes over one million clinical notes—discharge summaries and radiology reports—paired with structured admission-level features. Our study systematically progresses from sparse lexical representations (TF-IDF) used with classical models such as Logistic Regression, SVMs, Random Forests, and XGBoost, to contextual text encoders including BioClinicalBERT and Clinical-Longformer. We evaluate both Single-Task Learning (STL) and Multi-Task Learning variants, where auxiliary tasks such as in-hospital mortality and admission-type prediction are used to regularize and enrich patient-level representations.

Our findings reveal an interesting dynamic: while transformer models capture nuanced clinical semantics, a well-tuned Logistic Regression baseline using TF-IDF features remains the strongest overall performer (ROC-AUC: 0.703), underscoring the potency of lexical cues in clinical documentation. In contrast, BioClinicalBERT within an MTL framework achieves the most competitive deep-learning performance (ROC-AUC: 0.698), offering substantial recall gains beneficial for high-sensitivity clinical deployment. We further design a hybrid architecture that fuses transformer embeddings with structured data via gradient boosting, demonstrating the complementary strengths of neural and tabular modeling.

Overall, this work provides one of the most granular, model-level comparisons in clinical readmission prediction to date, highlighting when and why deep models help, where classical models remain surprisingly dominant, and how MTL strategies can be leveraged to improve representation learning across related clinical tasks.

1 Introduction

Hospital readmissions within 30 days of discharge remain a critical indicator of healthcare quality, resource utilization, and patient well-being. Unplanned readmissions contribute billions in additional annual expenditure and are often preventable through timely intervention. As healthcare systems continue shifting toward value-based care, the ability to accurately identify high-risk patients at discharge has become an essential clinical capability. Electronic Health Records (EHRs)—particularly the rich, narrative-style clinical notes written by physicians—offer an opportunity to build predictive models that anticipate adverse outcomes using both structured and unstructured information.

With the rapid advancement of Natural Language Processing (NLP), especially transformer architectures, there is growing interest in examining whether modern deep learning models can meaningfully outperform traditional approaches in clinical prediction tasks. Although clinical text contains nuanced temporal, contextual, and domain-specific language, the relative advantage of deep contextual models over simpler lexical methods is not yet fully understood for readmission prediction. Prior literature shows mixed evidence: some studies report improvements using pretrained biomedical transformers, while others find that classical methods such as Logistic Regression or gradient boosting remain highly competitive, often due to the strong signal present in surface-level text features.

This work aims to provide a comprehensive and

rigorous comparison across a spectrum of machine learning paradigms—ranging from classical linear models to state-of-the-art clinical transformers—under controlled experimental conditions. Unlike many previous studies that focus on a single modeling family, our approach is intentionally hierarchical and model-centric. We investigate:

- Classical ML with sparse lexical features, including Logistic Regression, SVMs, Random Forests, and XGBoost trained on TF-IDF vectors. These models provide strong baselines due to their robustness on high-dimensional, sparse clinical data.
- Single-Task Learning (STL) with transformer encoders, evaluating both frozen and fully fine-tuned versions of BioClinicalBERT and Clinical-Longformer. This allows us to isolate the contributions of pretraining, fine-tuning, and contextual representation quality.
- Multi-Task Learning (MTL) where readmission prediction is jointly optimized with auxiliary tasks such as mortality and admission-type prediction. MTL has the potential to improve representation learning by leveraging shared clinical semantics across related outcomes.
- Hybrid text–tabular modeling, where contextual embeddings are fused with structured EHR features and modeled through gradient boosting, combining neural representation learning with tabular model interpretability.

Through this multi-stage analysis, we aim to answer several key research questions:

- Do advanced transformer models significantly outperform classical models in predicting 30-day readmissions?
- Can auxiliary clinical tasks improve representation learning for readmission prediction via MTL?
- Which aspects of clinical documentation—lexical features, contextual semantics, or structured variables—offer the strongest predictive signal?
- What trade-offs exist between accuracy, recall, computational cost, and model complexity?

Using the MIMIC-IV dataset, we construct a large-scale, text-rich cohort with carefully engineered labels, note aggregation procedures, and imbalance mitigation strategies. Our experiments reveal that classical approaches, particularly Logistic Regression with TF-IDF features, continue to serve as remarkably strong baselines—surpassing the performance of both STL and MTL transformer models in overall ROC-AUC. However, transformer-based MTL approaches exhibit substantially higher recall, making them valuable in scenarios where missing at-risk patients poses significant clinical cost.

By presenting one of the most detailed end-to-end comparisons of modeling strategies for clinical readmission prediction, this work contributes to a deeper understanding of when sophisticated NLP architectures provide tangible benefits, how multi-task objectives can shape patient representations, and why classical lexical models continue to remain relevant in modern clinical machine learning pipelines.

2 Related Work

Research on predicting hospital readmissions spans multiple areas, including clinical NLP, structured EHR modeling, representation learning, and multi-task architectures. This section reviews progress in each direction and situates our contributions within the broader landscape.

2.1 Clinical Text Classification

Clinical text classification has long been a core task in healthcare NLP. Early approaches relied on sparse lexical representations such as bag-of-words and TF–IDF, which proved effective due to the formulaic, domain-specific structure of clinical documentation. Classical models (e.g., Logistic Regression, SVMs) achieved strong baseline performance in tasks such as ICD code assignment, mortality prediction, and adverse event detection.

Distributed word representations improved semantic modeling, with medical embeddings such as BioWordVec, FastText-Clinical, and embeddings trained over PubMed and MIMIC notes providing domain-aware vector spaces. However, these models lacked contextual understanding.

Transformer-based architectures marked a significant shift. Domain-specific variants such as BioBERT, ClinicalBERT, BioClinicalBERT, and BlueBERT demonstrated strong performance on named entity recognition, clinical classification,

concept extraction, and question answering. More recent models such as Clinical-Longformer and BigBird-RoBERTa address the challenge of long clinical documents by enabling 4k–16k token sequences, allowing full discharge summaries to be processed at once.

Despite these advances, several studies report that lexical signals can dominate performance in certain clinical tasks. Our work directly examines this tension by comparing sparse lexical baselines with contextual transformers for readmission prediction.

2.2 Readmission Prediction

Predicting early hospital readmission is a long-standing problem in clinical informatics. Traditional models rely heavily on structured EHR features such as demographics, comorbidities, diagnosis codes, and utilization patterns. Classical models—including Logistic Regression, Random Forests, and gradient boosting methods—remain widely used due to their robustness, interpretability, and ability to handle heterogeneous tabular inputs.

Incorporating clinical notes has emerged as a promising direction, motivated by the hypothesis that physician narratives capture psychosocial factors, disease severity, and discharge barriers not reflected in structured data. Early text-based models used bag-of-words, TF–IDF, and LDA-style topic models, demonstrating measurable improvements in performance.

More recent work employs pretrained clinical transformers such as ClinicalBERT and Longformer. These models capture nuanced semantic information and have been evaluated on readmission prediction, length of stay, ICU mortality, and risk stratification. However, the literature reports mixed results: while transformer models often capture richer structure, performance gains over lexical models remain inconsistent across datasets and tasks. Our work provides a controlled comparison of these approaches on MIMIC-IV.

2.3 Multi-Task Learning in Healthcare

Multi-Task Learning (MTL) has gained significant traction in healthcare due to its potential to improve generalization by sharing representations across related clinical outcomes. Prior research has explored joint prediction of mortality, length-of-stay, diagnosis codes, phenotypes, and next-visit forecasting, with results suggesting that shared encoders capture underlying patient trajectories more effectively than single-task models.

Transformer-based MTL architectures have been applied to resource-limited clinical datasets, leveraging auxiliary tasks for implicit regularization. Readmission, mortality, and admission-type prediction are highly correlated, making them strong candidates for joint optimization. Yet, comprehensive benchmarks comparing frozen versus fully fine-tuned clinical transformers in MTL settings remain limited. Our work addresses this by evaluating both BioClinicalBERT and Clinical-Longformer under multi-task configurations and quantifying their impact on readmission prediction.

2.4 Summary of Gaps Addressed

Although previous studies explore classical modeling, contextual transformers, and limited MTL settings, few directly compare multiple tiers of models under a unified pipeline. Specifically, existing work rarely:

- compares lexical baselines, classical ML, single-task transformers, multi-task transformers, and hybrid fusion models together,
- evaluates Longformer and BioClinicalBERT under identical experimental conditions,
- analyzes why classical models remain competitive for readmission prediction,
- or investigates how auxiliary tasks influence deep clinical representations.

We address these gaps through a systematic, model-centric analysis that clarifies the relative strengths and limitations of modern NLP techniques for clinical readmission prediction.

3 Data and Preprocessing

This section describes the dataset used in our study, the extraction pipeline developed for constructing the readmission cohort, and the preprocessing procedures for both structured and unstructured clinical data. Our goal was to create a text-rich, high-quality dataset that supports classical machine learning, transformer-based architectures, and multi-task learning models within a unified experimental framework.

3.1 Dataset

We use the MIMIC-IV v3.1 database, a large, publicly available collection of Electronic Health

Records (EHRs) covering over 380,000 hospital admissions from the Beth Israel Deaconess Medical Center. The dataset includes detailed structured information such as demographics, admission metadata, laboratory measurements, and outcomes, as well as millions of free-text clinical notes. All data were accessed through Google BigQuery (`physionet-data` project), which enabled efficient querying and scalable extraction of large text fields.

Our study focuses on three key MIMIC-IV components:

- **Admissions (hosp.admissions):** Provides structured patient-level metadata, including admission and discharge timestamps, insurance status, admission type, and in-hospital mortality.
- **Discharge Summaries (note.discharge):** Physician-authored summaries containing diagnoses, hospital course, medication changes, follow-up plans, and psychosocial considerations.
- **Radiology Notes (note.radiology):** Imaging reports that provide detailed descriptions of acute and chronic findings relevant to readmission risk.

Combining these sources allows our models to leverage both temporal admission patterns and rich clinical narratives.

3.2 Data Extraction Pipeline

We constructed an end-to-end extraction pipeline designed to produce a cleaned, aligned dataset suitable for readmission prediction and auxiliary task modeling. The pipeline consists of four main stages.

3.2.1 Readmission Label Generation

The primary binary label—30-day readmission—was derived using a window-based approach implemented via SQL analytic functions:

1. For each `subject_id`, rows were ordered by `admittime`.
2. We applied a `LEAD()` operation to obtain the next admission timestamp.
3. A 30-day window was computed using:

   $\Delta = \texttt{DATE\_DIFF(next\_admittime, dischtime)}$

4. A readmission was labeled positive if $0 \leq \Delta \leq 30$.

We additionally computed `days_to_readmission` as a continuous variable for exploratory analysis and error inspection.

3.2.2 Note Quality Filtering

Clinical notes vary widely in length and informativeness. To ensure usable training samples for both classical and transformer-based models, we applied length-based filtering:

- Discharge summaries shorter than **100 characters** were excluded.
- Radiology reports shorter than **50 characters** were removed.
- Admissions without any associated notes were discarded.
- Records containing missing or null text were filtered out.

This eliminated extremely brief or templated notes that provide minimal signal and would negatively affect text-based modeling.

3.2.3 Data Intersection and Cohort Construction

To create a coherent modeling cohort, we retained only those admissions that had:

1. a computed 30-day readmission label,
2. at least one valid discharge summary or radiology note,
3. complete admission metadata for auxiliary tasks.

This ensured that every sample contained both structured and unstructured data, allowing direct comparison across classical, transformer, MTL, and hybrid models.

3.3 Text Aggregation

Because multiple notes may be associated with a single hospital admission, we aggregated all available free-text documents at the `hadm_id` level. Notes were sorted chronologically and concatenated with delimiter tokens to preserve temporal ordering. This produced a single consolidated text record per admission.

Two types of aggregated text were created:

- **Discharge-only corpus:** used for models restricted to discharge summaries.

- **Combined corpus (discharge + radiology):** used for transformer-based models able to handle longer sequences.

The combined corpus offered richer clinical context but required long-sequence models such as Clinical-Longformer.

3.4 Handling Class Imbalance

Hospital readmission prediction is inherently imbalanced, with positive cases typically ranging from 10% to 20%. We employed multiple strategies to address this challenge:

- **Stratified Splitting:** Train/validation/test splits preserved the original class distribution.

- **Class Weighting:** Loss functions for classical and neural models used inverse-frequency weighting.

- **Undersampling for Balanced Batches:** Certain transformer experiments used balanced batches to stabilize training.

This ensured robust evaluation across model families without artificially inflating performance.

3.5 Final Dataset Statistics

Following extraction, filtering, and aggregation, we obtained a high-quality dataset consisting of:

- admissions with complete note coverage,

- structured variables (age, gender, insurance, admission type),

- readmission labels,

- auxiliary labels for mortality and admission type.

All processing was performed using a combination of BigQuery SQL and Python (`pandas`) to enable reproducibility and large-scale text handling. The resulting dataset forms the basis for all experiments described in Section 4.

4 Methodology

Our methodology follows a multi-stage modeling pipeline designed to evaluate a broad spectrum of machine learning approaches, ranging from sparse lexical baselines to transformer-based architectures and hybrid text–tabular models. This hierarchical design enables us to assess the incremental value added by contextual embeddings, multi-task objectives, and data fusion strategies.

We divide this section into four modeling phases: (1) classical baselines, (2) single-task transformers, (3) multi-task learning architectures, and (4) hybrid fusion models. Each phase builds on the previous one and is evaluated under identical preprocessing and train/test splits for fair comparison.

4.1 Phase 1: Classical ML Baselines

Classical machine learning models remain widely used for clinical text classification due to their efficiency, interpretability, and strong performance on high-dimensional sparse data. In this phase, we evaluate four baseline algorithms.

4.1.1 TF–IDF Feature Construction

Clinical notes were vectorized using TF–IDF with the following settings:

- Maximum vocabulary size: 50,000 unigrams.

- Minimum document frequency: 5.

- Sublinear term frequency scaling.

- Binary indicators disabled to preserve term frequencies.

This representation captures important lexical patterns while maintaining tractable sparsity for linear models.

4.1.2 Logistic Regression

We use ℓ_1- and ℓ_2-regularized Logistic Regression trained with the SAGA solver, chosen for its scalability to millions of sparse features. The model serves as a strong lexical baseline due to its stability, fast training, and interpretability via coefficient inspection.

4.1.3 Extended Benchmark Models

In addition to Logistic Regression, we evaluate:

Linear SVM (LinearSVC). Optimized using hinge loss with a linear kernel, suitable for high-dimensional text data.

Random Forest. Configured with depth constraints and limited splitting criteria to prevent overfitting on sparse TF–IDF vectors.

XGBoost. Utilized with the histogram-based tree method and `scale_pos_weight` to account for class imbalance. XGBoost provides non-linear decision boundaries that may capture interactions missed by linear models.

These baselines establish the performance level achievable using only lexical signals—a key comparison point for transformer-based models.

4.2 Phase 2: Single-Task Learning (STL)

This phase evaluates contextual transformers trained solely to predict 30-day hospital readmission. We explore two variants: frozen encoders and fully fine-tuned models.

4.2.1 Feature Extraction Baseline (Frozen Transformers)

We first evaluate Clinical-Longformer by freezing all encoder parameters and training only a lightweight classification head. This setup isolates the utility of pretrained clinical representations without task-specific adaptation. The classifier consists of:

- a linear projection of the `[CLS]` token,
- LayerNorm for stabilization,
- a softmax output layer.

Frozen models also serve as a computationally efficient baseline for large-scale experiments.

4.2.2 Full Fine-Tuning

For both BioClinicalBERT and Clinical-Longformer, we fine-tune all encoder layers using mixed-precision training. The classification head includes:

- a 768- or 1024-dimensional fully connected layer,
- GELU activation,
- dropout (rate = 0.1),
- a final softmax layer.

Training uses AdamW with linear warmup and early stopping. All sequences were truncated or padded to the maximum supported length of each model (512 tokens for BERT, 4096 for Longformer).

4.3 Phase 3: Multi-Task Learning (MTL)

To leverage correlated clinical outcomes, we design multi-task architectures that jointly predict:

1. 30-day readmission (primary task),
2. in-hospital mortality,
3. admission type (elective vs. emergency vs. urgent).

4.3.1 MTL Architecture

A shared transformer encoder processes the full clinical text. Three task-specific classification heads branch from the final hidden state:

$$h = \text{Encoder}(x), \qquad \hat{y}_i = \text{Head}_i(h)$$

The total loss is a weighted combination of cross-entropy terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{readmit}} + 0.5\,\mathcal{L}_{\text{mortality}} + 0.5\,\mathcal{L}_{\text{adm\_type}}$$

This weighting scheme reflects the centrality of readmission prediction while allowing auxiliary signals to regularize the encoder.

4.3.2 Frozen vs. Trainable Variants

We evaluate two MTL regimes:

- **Frozen encoder:** only task heads are trained.
- **Fully trainable encoder:** all transformer parameters are updated.

Frozen models test whether pretrained clinical semantics are sufficient; trainable models assess the benefit of deep adaptation.

4.3.3 Model-Specific Considerations

Clinical-Longformer enables long-sequence modeling of entire concatenated notes, whereas BioClinicalBERT relies on shorter summaries. This contrast allows us to evaluate whether including radiology + discharge notes improves performance when using long-context architectures.

4.4 Phase 4: Hybrid Architecture (Text + Structured Data)

While transformers excel at text representation, structured EHR features (e.g., age, gender, insurance, length of stay) often provide complementary information. To integrate both modalities, we propose a hybrid pipeline:

1. Extract the 768-dimensional `[CLS]` embedding from a fine-tuned BioClinicalBERT model.

2. Encode structured features via standard preprocessing (one-hot encoding, normalization).

3. Concatenate the text embedding with tabular features.

4. Train a LightGBM classifier on the combined representation.

This design allows gradient boosting to model non-linear interactions between text-derived embeddings and structured variables, while avoiding the computational overhead of jointly training multimodal neural networks.

4.5 Summary

Together, these four modeling phases provide a comprehensive spectrum of approaches for analyzing clinical readmission risk. This methodology allows us to evaluate not only the absolute performance of each model family but also their comparative advantages, limitations, and the diminishing or increasing returns associated with increased model complexity.

5 Experimental Setup

This section describes the computational environment, training configuration, hyperparameter choices, and evaluation metrics used for all experiments. Our goal is to ensure methodological consistency across classical baselines, transformer models, multi-task setups, and hybrid architectures, enabling robust and reproducible comparisons.

5.1 Hardware and Software Environment

All experiments were conducted using Python 3.11 and PyTorch 2.2 with HuggingFace Transformers. Depending on model size and training requirements, we used either local GPU hardware or cloud-based accelerators:

- **NVIDIA RTX 4080 Super (16 GB):** Used for classical ML, TF–IDF pipelines, Logistic Regression, SVMs, and frozen transformer baselines.

- **BigQuery SQL Engine:** Used for data extraction and label creation.

Mixed precision (FP16) training was enabled for all transformer models using `torch.cuda.amp`, significantly reducing memory usage without affecting performance.

5.2 Training Configuration

To ensure fairness, we standardize the training configuration across model families while allowing necessary architectural-specific choices.

5.2.1 Optimizer and Learning Rate Schedule

All deep learning models use the AdamW optimizer with:

- Weight decay: 0.01
- Initial learning rate: 2×10^{-5} (BERT models), 5×10^{-6} (Longformer)
- Batch size: 8–16 depending on sequence length
- Linear warmup: 10% of total steps
- Linear decay to zero after warmup

Classical models were trained with library defaults unless otherwise noted, with hyperparameters selected via small-scale grid search.

5.2.2 Memory Optimization

Due to the long-sequence nature of clinical notes, we employed:

- Gradient checkpointing for Clinical-Longformer (reducing memory by $\sim$40%),
- Chunked loading of large text fields,
- Sequence truncation to model-specific limits (512 for BERT, 4096 for Longformer),
- Automatic mixed precision (AMP).

These optimizations enabled full fine-tuning without exceeding GPU memory limits.

5.2.3 Early Stopping and Validation

For all neural models:

- A 10% validation split was used from the training set.
- Early stopping patience was set to 3 validation epochs.
- Best checkpoints were selected based on validation ROC-AUC.

This prevented overfitting, particularly in MTL models with multiple loss terms.

5.3 Model-Specific Hyperparameters

Each model family required additional configuration:

Classical Models.

- Logistic Regression: SAGA solver, $C \in \{0.1, 1.0, 10.0\}$ evaluated.
- Linear SVM: hinge loss, $C = 1.0$.
- Random Forest: 200 trees, max depth = 25.
- XGBoost: learning rate = 0.1, max depth = 6, `scale_pos_weight` tuned based on imbalance.

STL Transformers.

- Maximum sequence length: 512 (BioClinicalBERT), 4096 (Longformer).
- Dropout rate: 0.1.
- Classifier hidden dimension: 1024 (Longformer), 768 (BERT).

MTL Transformers.

- Three separate classification heads with independent dropout layers.
- Loss weighting: $\mathcal{L} = \mathcal{L}_r + 0.5\mathcal{L}_m + 0.5\mathcal{L}_t$.
- Batch size reduced to fit the added memory cost.

Hybrid Model (Text + Tabular).

- `[CLS]` embeddings extracted from fine-tuned BioClinicalBERT.
- LightGBM with 400 trees, learning rate 0.05.
- Early stopping on a 15% validation split.

5.4 Evaluation Metrics

Given the imbalanced nature of 30-day readmissions, we prioritize metrics that remain informative under skewed class distributions.

5.4.1 Primary Metric: ROC-AUC

We use ROC-AUC as the primary metric across all models because it:

- captures global ranking performance,
- is robust to class imbalance,
- enables direct comparability between model families.

5.4.2 Secondary Metrics

To complement ROC-AUC, we report:

- **Precision** (positive predictive value),
- **Recall** (sensitivity), especially important for minimizing missed readmissions,
- **F1-score** as the harmonic mean of precision and recall,
- **Accuracy** for completeness (not used for comparison).

5.4.3 Balanced vs. Imbalanced Evaluation

We evaluate on:

- the natural (imbalanced) test set, and
- a class-balanced test sample (for STL transformer analysis).

This dual evaluation reveals whether improvements arise from better ranking performance or changes in class sensitivity.

5.5 Reproducibility

To ensure reproducibility:

- all random seeds were fixed (PyTorch, NumPy, Python),
- full code and hyperparameter settings were logged,
- deterministic cuDNN operations were enabled when feasible.

This setup guarantees consistent results across reruns and supports future extensions of this work.

6 Results

This section presents quantitative results for all model families studied in our hierarchical pipeline. We begin by reporting performance for classical TF–IDF-based models, followed by single-task transformers, multi-task architectures, and the hybrid fusion model. We evaluate each model using the metrics outlined in Section **??**, with ROC-AUC as the primary metric. All results are averaged across three random seeds.

Model	AUC	Recall	F1
LR (Full)	0.7003	0.6051	0.4260
LR (Balanced)	**0.7031**	0.6275	**0.6400**
XGBoost	0.6898	0.5524	0.4143
Linear SVM	0.6831	0.5945	0.4123
Random Forest	0.6729	0.4425	0.3924

Table 1: Classical ML baseline performance using TF–IDF features.

6.1 Classical ML Performance

Table 1 summarizes the performance of classical models trained on TF–IDF features. Logistic Regression remains the strongest baseline, achieving the highest ROC-AUC across all classical models.

These results highlight several trends:

- **Lexical dominance:** Simple unigram features capture a substantial portion of the predictive signal present in clinical notes.
- **Model sensitivity:** Logistic Regression outperforms more complex ensemble methods, suggesting that linear decision boundaries are sufficient for this task.
- **Impact of class balancing:** The balanced training setup increased recall and F1 significantly without degrading AUC.

Overall, this phase establishes a strong baseline against which transformer models must compete.

6.2 Deep Learning Performance

Transformer-based models were evaluated under both single-task learning (STL) and multi-task learning (MTL) settings. Table 2 reports performance across all configurations.

Several observations emerge:

- **MTL improves AUC:** BioClinicalBERT under MTL achieves the best deep-learning performance, nearly matching Logistic Regression.
- **Recall-precision trade-off:** Clinical-Longformer MTL yields extremely high recall (80.5%), making it suitable for applications prioritizing sensitivity.
- **Pretraining limitations:** Frozen STL transformers perform significantly worse, confirming that clinical pretraining alone is insufficient without fine-tuning.
- **Sequence length considerations:** Despite its ability to model long text, Clinical-Longformer does not outperform BioClinicalBERT, suggesting diminishing returns from modeling full note context.

These results demonstrate that transformer-based models bring clear benefits in recall and representation learning, but do not always surpass classical baselines in overall discrimination performance.

6.3 Hybrid Fusion Results

The hybrid architecture combining BioClinicalBERT embeddings with LightGBM achieves moderate improvements over STL transformers but does not reach the performance of the MTL models. This suggests that while structured data adds complementary information, text signals remain dominant for this specific task.

The hybrid approach is valuable for interpretability and integration into clinical workflows, though its predictive power is limited by the quality of the underlying text embeddings.

6.4 Key Findings

Across all experiments, several themes emerged:

1. **MTL provides meaningful gains.** BioClinicalBERT with multi-task learning consistently improves ROC-AUC relative to STL, supporting the hypothesis that related clinical outcomes share useful representations.
2. **Classical models remain competitive.** Despite their simplicity, TF–IDF + Logistic Regression matches or exceeds transformer models in ROC-AUC. This reinforces the strength of lexical features in clinical documentation.
3. **Recall vs. precision trade-offs.** MTL-Longformer shows extremely high recall but low precision, making it suitable when minimizing false negatives is essential.
4. **Frozen transformers underperform.** Lack of fine-tuning results in substantial performance degradation, indicating that task-specific adaptation is crucial even for clinically pretrained encoders.
5. **Hybrid fusion is promising but limited.** Combining structured data with text embeddings yields modest benefits but does not surpass MTL models.

Architecture	Strategy	AUC	Recall	F1
BioClinicalBERT	MTL	**0.6984**	0.566	0.420
BioClinicalBERT	STL†	0.6666	0.586	**0.608**
Clinical-Longformer	MTL	0.6770	**0.805**	0.393
Clinical-Longformer	MTL (Frozen)	0.5367	0.981	0.351
Clinical-Longformer	STL	0.6596	0.422	0.359
Clinical-Longformer	STL (Frozen)	0.5899	0.025	0.046
Hybrid (BERT+LGBM)	Fusion	0.6602	0.560	0.387

Table 2: Comparison of deep learning models. †Evaluated on a balanced test set.

These findings provide a nuanced understanding of the benefits and limitations of modern NLP methods for clinical readmission prediction and inform the design of future hybrid and multitask clinical prediction systems.

7 Discussion

Our results reveal several important insights regarding the comparative performance of classical machine learning models, contextual transformers, multi-task architectures, and hybrid text–tabular systems for hospital readmission prediction. This section discusses these findings, interprets their implications for clinical NLP research, and highlights the broader lessons learned from our systematic evaluation.

7.1 Why Classical Models Remain Competitive

One of the most striking results from our study is the strong performance of Logistic Regression with TF–IDF representations. Despite the availability of large pretrained transformers and sophisticated MTL architectures, the classical baseline achieves the highest overall ROC-AUC (0.703), surpassing all transformer-based models.

There are several factors that explain this trend:

- **Lexical dominance in clinical notes.** Clinical documentation often contains highly predictive keywords (e.g., “hospice”, “palliative”, “metastatic”, “acute renal”, “end-stage”) whose presence or absence drives risk stratification. TF–IDF captures these signals effectively without requiring deep semantic modeling.
- **Formulaic structure of discharge summaries.** Many discharge summaries follow standardized templates. As a result, key phrases, sections, and terminology recur consistently, making sparse lexical models particularly effective.
- **Sample efficiency.** Transformers require large amounts of task-specific training data to fully exploit their contextual capabilities. TF–IDF models, in contrast, generalize well even with moderate-sized datasets.
- **Regularization benefits.** Linear models with ℓ_1/ℓ_2 penalties generalize strongly on high-dimensional sparse data, avoiding overfitting even without heavy architectural constraints.

Taken together, these factors highlight that sophisticated modeling is not always synonymous with superior performance in clinical prediction tasks.

7.2 Value and Limitations of Transformer Models

Although transformer-based models do not surpass classical methods in ROC-AUC, they offer substantial improvements in other dimensions:

- **Higher recall.** Clinical-Longformer MTL-Frozen achieves the highest recall in our study (0.98), suggesting that contextual models may be better at identifying subtle patterns in patient deterioration or complications.
- **Better representation learning.** BioClinicalBERT demonstrates consistent improvements under multi-task learning, implying that shared clinical semantics across readmission, mortality, and admission type help shape a more informative representation space.
- **Stronger performance in balanced evaluation.** BERT-based STL models outperform classical methods on a balanced subset of the test data, indicating improved discrimination when class frequencies are normalized.

However, limitations remain:

- **Computational overhead.** Transformers require substantial GPU resources, long training

times, and careful memory management, making them less practical for routine clinical deployment without specialized infrastructure.

- **Sequence truncation effects.** For BERT-based models, the 512-token limit forces aggressive truncation of long clinical notes. In many cases, the most predictive information appears early in the note, which may reduce the relative advantage of long-context transformers.

- **Sensitivity to noise.** Transformers can inadvertently overfit to stylistic or institution-specific patterns in documentation, particularly when the dataset size is limited relative to model capacity.

These findings highlight that transformers are most useful when recall is prioritized or when auxiliary clinical tasks can meaningfully augment representation learning.

7.3 Effectiveness of Multi-Task Learning

The multi-task architecture provides notable benefits over single-task models. For both BioClinicalBERT and Clinical-Longformer, MTL improves ROC-AUC and stabilizes learning, suggesting that auxiliary clinical outcomes serve as effective inductive biases:

- **Shared patient trajectory signals.** Mortality and admission type share contextual determinants with readmission, encouraging the model to learn general clinical concepts such as disease severity, resource use, and progression patterns.

- **Regularization through shared features.** By training on multiple tasks, the model avoids overfitting to task-specific artifacts in discharge summaries.

- **Improved sensitivity.** MTL-Longformer demonstrates extremely high recall, indicating that auxiliary tasks help the model identify high-risk patients more reliably.

However, MTL does not consistently outperform classical baselines in terms of ROC-AUC, suggesting that auxiliary-task gains may be strongest when operations prioritize sensitivity or early detection rather than overall ranking performance.

7.4 Hybrid Fusion: Advantages and Limitations

The hybrid model combining BioClinicalBERT embeddings with structured data via LightGBM provides moderate performance benefits but does not reach the levels achieved by MTL. This suggests:

- **Structured data contributes complementary but modest value.**

- **Text-based signals dominate readmission risk.**

- **Interacting multimodal features may require deeper fusion** (e.g., joint neural tabular–text models) to fully exploit their potential.

The hybrid approach remains practical for deployment in clinical workflow settings where interpretability and modularity are valued.

7.5 Implications for Clinical Deployment

From a systems and workflow perspective, our results suggest the following considerations:

- **Classical models are strong deployment candidates** due to low latency, low cost, and competitive performance.

- **Transformer models may be preferable in high-risk settings** where missing a readmission has critical consequences.

- **MTL architectures show promise for generalized clinical risk models** that predict multiple outcomes simultaneously.

Thus, the "best" model depends on the clinical objective: minimizing false negatives, maximizing ranking performance, or minimizing infrastructure cost.

7.6 Broader Reflections

Finally, our study reinforces several broader themes relevant to clinical machine learning:

- **Bigger models do not guarantee better performance.**

- **Clinical data often reward lexical cues over deep semantics.**

- **Auxiliary tasks provide strong inductive bias.**

- **Interpretability and efficiency remain critical for adoption.**

These insights will help guide future work on model selection, architectural design, and clinical adaptation of NLP systems for health applications.

8 Conclusion and Future Work

In this work, we conducted a comprehensive and systematic evaluation of machine learning and deep learning approaches for 30-day hospital readmission prediction using clinical notes and structured EHR data from MIMIC-IV. Our hierarchical methodology enabled us to examine the full modeling spectrum, from lightweight classical methods to large transformer architectures and multi-task learning frameworks. The results reveal a nuanced landscape in which different modeling families excel along different dimensions of performance, complexity, and clinical utility.

8.1 Summary of Contributions

Our study provides several key contributions:

- We demonstrated that **classical text-based models remain remarkably strong baselines**, with Logistic Regression + TF–IDF achieving the highest ROC-AUC across all models evaluated.
- We showed that **transformer models provide substantial gains in recall**, particularly under multi-task learning, making them attractive in scenarios where early identification of at-risk patients takes priority.
- We evaluated **multiple transformer architectures** (BioClinicalBERT and Clinical-Longformer) under both single-task and multi-task settings, providing one of the clearest comparative analyses in the clinical NLP literature.
- We developed and assessed a **hybrid text–tabular fusion model**, highlighting the complementarities and limitations of multimodal modeling.
- We provided a **detailed, reproducible modeling pipeline**, including preprocessing, labeling, and data curation procedures, that may serve as a template for future research using MIMIC-IV and similar datasets.

Together, these contributions offer a clearer understanding of where different modeling paradigms excel and underscore the importance of establishing strong, transparent baselines in clinical prediction tasks.

8.2 Implications for Clinical NLP Research

Our findings challenge the common assumption that larger, more complex models always yield superior predictive performance. Instead, we observe:

- Classical lexical models remain competitive due to the structured, keyword-driven nature of clinical documentation.
- Transformer-based models benefit significantly from multi-task learning, which reinforces shared clinical semantics.
- Sequence length alone does not guarantee improvement, as demonstrated by the limited gains from long-context models like Longformer.

These insights suggest that future clinical NLP research should carefully consider model selection, task formulation, and the intrinsic properties of clinical free text before adopting computationally expensive architectures.

8.3 Future Work

Several directions emerge from our findings:

1. Advanced Hyperparameter Optimization. Future research may apply Bayesian optimization or population-based training to tune MTL loss weights, learning rates, and transformer-specific hyperparameters, potentially unlocking further gains.

2. Ensemble and Stacking Methods. Ensembling classical and transformer models may combine lexical and semantic signals more effectively than hybrid fusion alone. Model stacking or meta-learners could also help improve calibration.

3. Explainability and Interpretability. Applying SHAP, Integrated Gradients, or attention-based explanation methods could help elucidate the textual cues driving predictions and increase clinical trust.

4. Incorporation of Structured Time-Series Data. Vital signs, lab trajectories, and medication timelines remain unexplored in our setting. Combining temporal structured data with text may provide substantial gains.

5. Exploration of Prompt-Based or Adapter-Based Large Language Models. Parameter-efficient fine-tuning (PEFT), domain adapters, or prompt tuning using modern LLMs may offer improved performance at significantly lower computational cost.

6. Generalization Across Institutions. Future studies should evaluate model portability to other hospital systems and clinical settings to assess robustness and fairness.

8.4 Final Remarks

Overall, our results highlight that **model complexity must be balanced against practical utility** in clinical machine learning. While transformers and multi-task methods provide valuable improvements in specific dimensions, classical models remain strong, interpretable, and computationally efficient tools for real-world deployment. Continued exploration of hybrid, multitask, and interpretable modeling approaches will be essential as healthcare systems increasingly integrate NLP-driven risk prediction workflows into clinical practice.

9 Limitations

While our study provides a comprehensive evaluation of multiple modeling approaches for 30-day readmission prediction, several limitations must be acknowledged. These limitations inform the appropriate interpretation of our results and highlight opportunities for future improvement.

9.1 Dataset and Generalizability

Our work is based exclusively on the MIMIC-IV dataset, which captures ICU and hospital admissions from a single academic medical center. Although MIMIC-IV is widely used in clinical machine learning research, it may not reflect the documentation styles, patient demographics, operational workflows, or institutional practices of other hospitals. As a result:

- Model generalizability to non-ICU or community hospital settings remains uncertain.
- Institution-specific clinical language and templating practices may bias lexical and transformer-based models.
- External validation across multiple healthcare systems is needed to confirm robustness.

Thus, caution should be exercised in interpreting our findings as universally applicable across all clinical environments.

9.2 Scope of Clinical Notes

Although we incorporate both discharge summaries and radiology notes, our study does not include other relevant textual sources such as nursing notes, progress notes, or medication instructions. These additional modalities may contain complementary information, including:

- day-to-day changes in clinical status,
- psychosocial considerations and discharge barriers,
- medication adherence challenges,
- social determinants of health.

Limiting text sources may constrain the full predictive potential of transformer models, especially those designed for long-sequence analysis.

9.3 Limited Structured Feature Set

Our structured data features are intentionally restricted to demographic and admission-level variables to maintain a clean focus on text-based modeling. However, important predictors such as:

- laboratory trends,
- comorbidity indices,
- medication history,
- prior healthcare utilization,
- vital sign trajectories,

were not included. Their absence likely limits the performance of both transformer and hybrid models, especially those that rely on multimodal context.

9.4 Sequence Truncation and Model Constraints

Despite evaluating Clinical-Longformer, several constraints remain:

- BioClinicalBERT is limited to 512 tokens, necessitating truncation of long notes.
- Even Longformer (4096 tokens) cannot always accommodate extremely long aggregated admission records.

- Truncation may disproportionately remove clinically relevant content from late sections such as follow-up instructions or discharge medications.

Future work with hierarchical transformers or retrieval-augmented architectures may mitigate these issues.

9.5 Model Complexity and Computational Cost

Transformer training—particularly under MTL—requires significant GPU memory, extended training time, and careful engineering. These constraints limit:

- the extent of hyperparameter search,
- the ability to test larger architectures or longer sequences,
- the feasibility of deploying such models in low-resource clinical settings.

Classical models are trivial to train by comparison, reinforcing their appeal for real-world use despite their simplicity.

9.6 MTL Task Selection

Our multi-task setup uses mortality and admission type as auxiliary tasks. While these are clinically meaningful and correlated with readmission, they represent only a subset of possible auxiliary signals. Other related tasks such as:

- length of stay prediction,
- diagnosis code prediction,
- severity scoring,
- readmission at 7 or 90 days,

may yield different or stronger benefits. Thus, our MTL conclusions should be interpreted within the context of this specific task combination.

9.7 Evaluation Limitations

Although we evaluate using ROC-AUC, F1, precision, and recall, other evaluation dimensions remain unexplored, including:

- model calibration (e.g., Brier score),
- decision-curve and cost-sensitive analyses,
- clinical utility under operational constraints.

These analyses are crucial for deploying predictive models in clinical environments and assessing actual patient-level impact.

9.8 Interpretability Constraints

Finally, many transformer models behave as "black boxes," making their decision-making processes difficult to interpret for clinicians. While classical models offer transparent coefficients, transformer interpretability techniques (attention maps, gradient-based methods, SHAP) were not explored in this study. This limits the clinical trustworthiness of the more complex models.

9.9 Summary

Overall, these limitations highlight the need for broader datasets, richer modalities, deeper hyperparameter exploration, and stronger interpretability efforts. Addressing these gaps will be essential for improving the generalizability, fairness, and clinical relevance of future readmission prediction models.