MedPredict

Hospital Readmission Prediction System with LLM

Team 06 :

Jaya Lakshmi Gunji

Naveen Yadav Gongati

Neha Dabeeru

Priyanka Akula

Somna Sattoor

Table of Contents

01 Introducing MedPredict!!

Introduction:

This system uses large language models (LLMs) to analyze medical data and predict at-risk patients, enabling early interventions, reducing readmissions, and improving patient outcomes while optimizing resource use.

Motivation:

With rising healthcare costs and operational pressures, this solution addresses the need for early identification of at-risk patients, helping reduce readmissions, prevent complications, and enhance care efficiency.

Objectives:

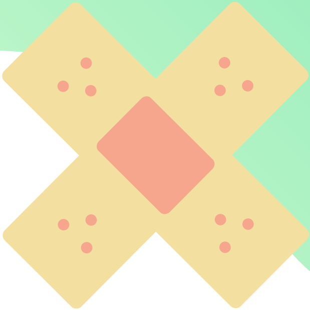
- Predict at-risk patients using LLMs.
- Provide actionable insights for timely interventions.
- Reduce readmissions and healthcare costs.
- Improve patient outcomes and resource management.

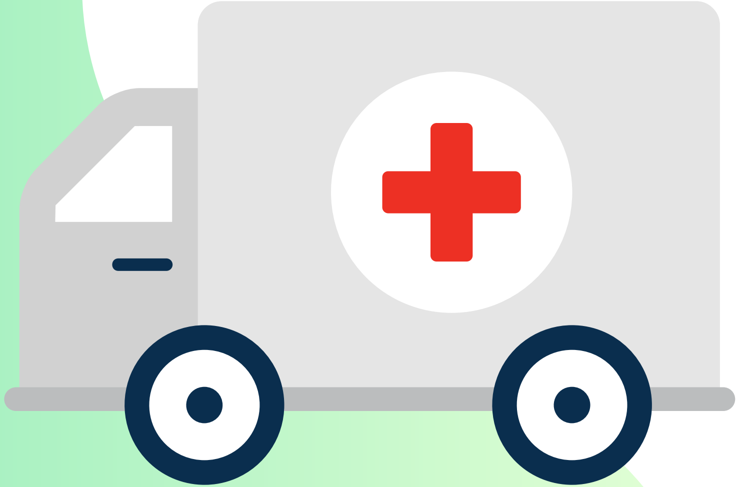
Needs:

Access to medical data, computational resources, and collaboration with healthcare providers for real-world implementation and continuous model improvement

02 Literature and Technology Survey

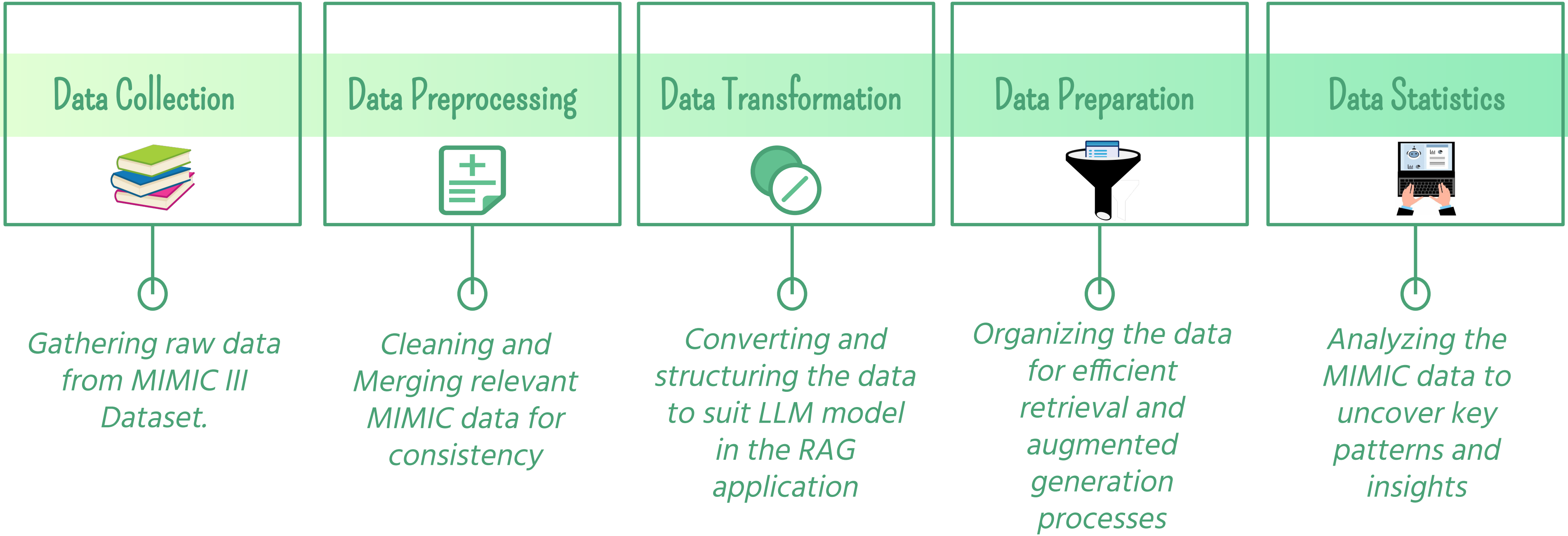
Author	Technical Approach	Models	Evaluation Metrics	Result	Conclusion
Pawar et al. (2022)	Supervised Fine-Tuning, Multimodal Integration	Clinical BERT	Accuracy, F1 Score	Improved prediction accuracy for 30-day mortality in COVID-19 patients	Demonstrates the benefits of multimodal models and fine-tuning with structured data to enhance Clinical BERT's accuracy in predicting hospital readmissions.
Shahandashti et al. (2024)	Eliminative Argumentation, CoT prompting, predicate-based rules	GPT-4 Turbo	Efficacy in defeater detection	Reasonable efficacy in detecting defeaters in assurance cases	Highlights the potential of GPT-4 Turbo in safety and reliability assurance for critical sectors.
Chai et al. (2023)	Noise reduction with Shared Labels and Dynamic Splicing	XLNet-CRF	F1 score improvement	Achieved state-of-the-art performance on 7	Proposes effective noise reduction models for BioNER systems.
Li et al. (2022)	Sparse attention mechanisms for long clinical sequences	Clinical-Longformer, Clinical-BigBird	Named entity recognition, QA, Document classification	Superior performance compared to ClinicalBERT, especially on extended sequences	Demonstrates capability to handle long-term dependencies in clinical text, improving decision-making.
Li et al. (2024)	Benchmarking LLMs for evidence-based medicine	PubMedBERT	PICO extraction performance, Biomedical	13% boost in PICO extraction	Highlights the strong potential of PubMedBERT but underlines the need for human oversight.





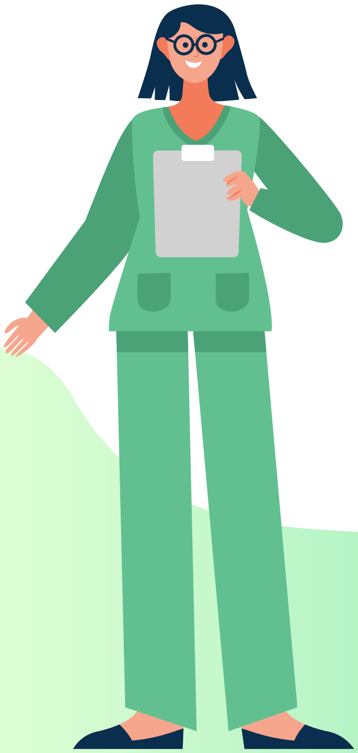
03

Data Engineering


Data Collection
Gathering raw data from MIMIC III Dataset.
Data Preprocessing
Cleaning and Merging relevant MIMIC data for consistency
Data Transformation
Converting and structuring the data to suit LLM model in the RAG application
Data Preparation
Organizing the data for efficient retrieval and augmented generation processes
Data Statistics
Analyzing the MIMIC data to uncover key patterns and insights

Data Collection

The MIMIC dataset contains several CSV files, but for this project, we focused on 6 key files:



- **PRESCRIPTIONS.csv**: Medication details prescribed during hospital stays.
- **ADMISSIONS.csv**: Admission and discharge data for patient tracking.
- **D_ICD_DIAGNOSES.csv**: ICD codes for medical diagnoses classification.
- **DIAGNOSES_ICD.csv**: Patient diagnosis data linked to ICD codes.
- **ICUSTAYS.csv**: ICU stay records, including duration and outcomes.
- **PATIENTS.csv**: Demographic details (age, gender) for patient analysis.

```
Shape of ADMISSIONS.csv: (129, 19)
Shape of DIAGNOSES_ICD.csv: (1761, 5)
Shape of D_ICD_DIAGNOSES.csv: (14567, 4)
Shape of ICUSTAYS.csv: (136, 12)
Shape of PATIENTS.csv: (100, 8)
Shape of PRESCRIPTIONS.csv: (10398, 19)
```

Data Preprocessing and Transformation:

Missing Values: Imputed nulls; categorical fields with mode, numerical with 0.

Data Merge & Cleanup: Merged 6 CSVs, removed irrelevant columns (e.g., religion, ethnicity).

Date Adjustment: Normalized date columns to ensure valid range (1980-2023).

corrected_year
2014
1985
1985
1999
2013

years not in range for admittime: [1976 1975 2045 2040 1979 2035 2044 1957 1977 2039 2038 2042 1974 2030
1978 1955 1970 1967 1965 2048 1973 2043 1968 2036 2034 2049 2050 2051
2052 1962 1952 1969 1971 1960 2029 2026 1956 1954 2025 2028]

	0
row_id	0
subject_id	0
hadm_id	0
admittime	0
admission_type	0
admission_location	0
discharge_location	0
insurance	0
diagnosis	0
emergency_stay_duration	0
short_title	0
icu_type	0
icu_duration	0
gender	0
drug	0
drug_type	0
route	0
prescription_duration	0
total_dosage	0
readmitted_within_30_days	0

Data Preprocessing and Transformation:

Text Normalization: Standardized text fields by converting them to lowercase, ensuring consistency for vectorization.

	original_admission_type	admission_type	original_admission_location	admission_location	original_discharge_location	discharge_location	original_diagnosis	diagnosis
0	EMERGENCY	emergency	EMERGENCY ROOM ADMIT	emergency room admit	HOME HEALTH CARE	home health care	SEPSIS	sepsis
1	EMERGENCY	emergency	TRANSFER FROM HOSP/EXTRAM	transfer from hosp/extram	DEAD/EXPIRED	dead/expired	HEPATITIS B	hepatitis b
2	EMERGENCY	emergency	TRANSFER FROM HOSP/EXTRAM	transfer from hosp/extram	DEAD/EXPIRED	dead/expired	SEPSIS	sepsis

readmitted_within_30_days	emergency_stay_duration	icu_duration
0	6.283333	1.6325
0	0.000000	13.8507
0	0.000000	2.6499

Feature Engineering:

Emergency Stay Duration: Time difference between edregtime and edouttime.

Readmitted Within 30 Days: Binary feature for LLM to classify patient readmissions within 30 days.

Prescription & ICU Durations: Duration based on start and end times for prescriptions and ICU stays.

Patient Age: Calculated by subtracting dob from admittime.

Data Preprocessing and Transformation:

Handling Duplicates: Check duplicate rows based on *subject_id* and *hadm_id* to ensure data integrity.

```
Duplicated rows: 43861
```

Class Imbalance: For predicting "readmitted_within_30_days," we tackled the class imbalance by generating synthetic data using random functions, enhancing our model's performance on this specific prediction.

```
readmitted_within_30_days
0    86.821705
1    13.178295
Name: proportion, dtype: float64
```

```
Number of patients with high risk of readmission: 500
Number of patients with low risk of readmission: 500
```

Data Preparation:

Generating Clinical Summaries: Created personalized clinical summaries for each patient, integrating relevant details like diagnosis, age, and prescriptions. These summaries can serve as a contextual reference when retrieving information for the LLM

subject_id	hadm_id	clinical_notes
10006	142345	The patient, identified as a 64-year-old female (Patient ID 10006, Hospital Admission ID 142345), was admitted to our facility on 2014-10-23 under the classification of emergency. The patient presented to the hospital via emergency room admit, indicating a possible referral from another healthcare facility or transfer from emergency services. After completing the necessary course of care, the patient was eventually discharged to home health care. Their hospital stay was covered under the Medicare insurance plan, which facilitated comprehensive treatment. Upon admission, the primary diagnosis was determined to be sepsis (also referred to as Sepsis). Given the severity of the patient's symptoms, immediate medical attention was required. The patient initially spent approximately 6.283333333333333 hours in the emergency department, during which time they were stabilized and subjected to a range of diagnostic tests. Due to the critical nature of their condition, it was decided that further intensive monitoring and intervention were necessary. The patient was transferred to the carevue ICU, where they remained for 1.6325 days. During their ICU stay, the patient was closely monitored and received the appropriate medical interventions tailored to their condition. Notably, the patient was administered sodium chloride 0.9% flush, categorized as a MAIN medication, to manage their condition effectively. The medication was delivered via IV, ensuring optimal therapeutic benefit. The prescribed regimen lasted for 9 days, with a total administered dosage amounting to 3.0 units. The choice and administration of this medication were crucial in stabilizing the patient's condition and promoting recovery. The patient remained in the hospital for a total of 9 days, during which time their progress was closely monitored. Upon discharge, the patient was deemed stable and capable of continuing recovery at home. Importantly, based on follow-up information, it was noted that the patient was no readmitted within 30 days of discharge. This suggests that the initial treatment plan was effective and that the patient adhered well to post-discharge care protocols. Overall, the patient responded positively to the interventions provided during their stay, and their eventual discharge was a result of significant clinical improvement. Ongoing outpatient follow-up will ensure that any further complications are addressed promptly.

Data Preparation:

Conversational Clinical Notes: We are utilizing a JSON file containing conversational clinical notes to train the LLM, enabling it to better understand and generate context-specific medical insights.

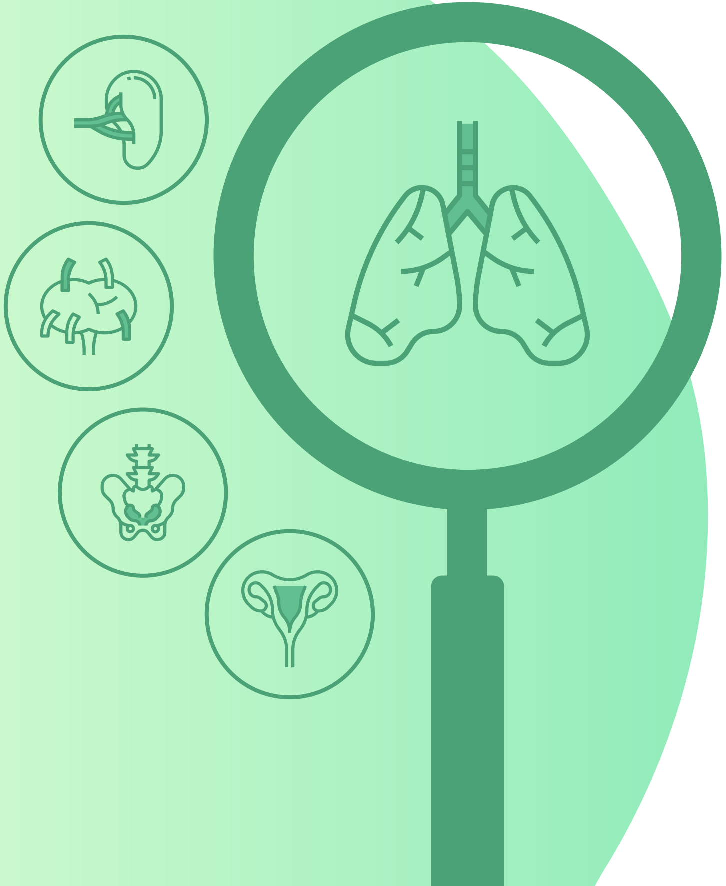

```
        "id": "12269",
        "dialogue": [
            {
                "role": "model",
                "content": "Hello, I am here to help assess your readmission risk. Please provide your gender, diagnosis, admission type, discharge location, insurance, number of prescriptions, ICU duration, ICU type, route of medication administration, drug type, specific drugs taken, and emergency stay duration."
            },
            {
                "role": "patient",
                "content": "I am a female with the diagnosis of humeral fracture admitted as a emergency case. Discharged to SNF, covered by Medicare insurance. I received 105 prescriptions, stayed 2.1436 days in the ICU as a carevue patient, medication was administered via SC as MAIN and the drugs included hydromorphone. My emergency room stay lasted 457.0 days."
            },
            {
                "role": "model",
                "content": "Thank you for the information. Based on the details provided, there is a low risk of readmission within 30 days."
            }
        ]
    },
```

Fine-Tuned Enhanced Dataset

{"patient_id":7,"age":80,"gender":"male","marital_status":"divorced","diagnosis":"heart failure","number_of_prior_admissions":1,"insurance":"private","admission_type":"elective","admission_location":"transfer from skilled nursing facility","discharge_location":"home with IV services","number_of_admitted_days":"long stay of 10 days","mean_lab_results":{"creatinine":"2.26 mg\/dL","sodium":"139.85 mmol\/L","glucose":"110.97 mg\/dL","urea":"12.9 mg\/dL"},"last_lab_results":{"creatinine":"2.46 mg\/dL","sodium":"110.25 mmol\/L","glucose":"93.05 mg\/dL","urea":"0.0 mg\/dL"},"mean_recorded_vitals":{"heart_rate":"86.14 bpm","blood_pressure":"112.24 mmHg","temperature":"37.46 \u00b0C","respiratory_rate":"20.24 bpm"},"last_recorded_vitals":{"heart_rate":"76.02 bpm","blood_pressure":"129.77 mmHg","temperature":"36.29 \u00b0C","respiratory_rate":"18.91 bpm"},"number_of_medications":"moderate medication diversity of 40 unique medications","number_of_days_in_emergency_room":"moderate duration of 4.18 hours","readmission_status":"no","enhanced_summary":"Based on the information provided, the 80-year-old male patient with heart failure was admitted to the hospital from a skilled nursing facility for elective care. He had a long stay of 10 days in the hospital and was discharged home with IV services. His latest lab results showed elevated creatinine levels and low sodium levels. His vital signs were within normal range. The patient was on a moderate diversity of 40 unique medications. He spent 4.18 hours in the emergency room but did not readmit to the hospital within 30 days."}
{"patient_id":8,"age":68,"gender":"male","marital_status":"married","diagnosis":"pneumonia","number_of_prior_admissions":0,"insurance":"medicaid","admission_type":"elective","admission_location":"emergency room","discharge_location":"home","number_of_admitted_days":"medium stay of 6 days","mean_lab_results":{"creatinine":"2.45 mg\/dL","sodium":"98.0 mmol\/L","glucose":"105.94 mg\/dL","urea":"14.9 mg\/dL"},"last_lab_results":{"creatinine":"2.49 mg\/dL","sodium":"164.35 mmol\/L","glucose":"102.29 mg\/dL","urea":"3.8 mg\/dL"},"mean_recorded_vitals":{"heart_rate":"84.69 bpm","blood_pressure":"124.37 mmHg","temperature":"37.45 \u00b0C","respiratory_rate":"19.44 bpm"},"last_recorded_vitals":{"heart_rate":"74.96 bpm","blood_pressure":"130.05 mmHg","temperature":"36.3 \u00b0C","respiratory_rate":"17.05 bpm"},"number_of_medications":"low medication diversity of 20 unique medications","number_of_days_in_emergency_room":"moderate duration of 5.40 hours","readmission_status":"no","enhanced_summary":"Additional Notes: The patient responded well to treatment for pneumonia and was discharged home in stable condition. The patient is advised to follow up with their primary care physician for further monitoring and management of their health."}

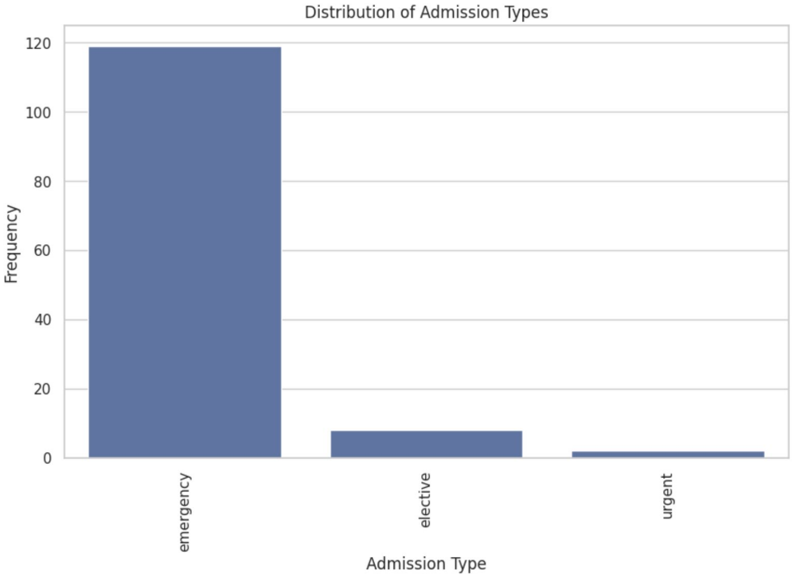
Data Statistics

Stage	Steps	Applicable Statistics
Raw Data	MIMIC III Dataset	60 GB
Sampled Data	Manual Sampling of Excel Data	approximately 2.93 MB
Data pre-processing	Null Value Handling	14,171 instances
	Inconsistent Data Handling	39 instances in ICU and Patient CSV
	Detect Outliers	20 instances removed
	Handling Duplicates	43861 instances
Data transformation	Text Normalizations	Lower case to maintain consistency among admission_types, admission_location
	Feature Engineering	Created new features like emergency_duration, readmitted_30_days, ICU_duration, etc
	Class Imbalance	Readmitted_30_days with 86:13 ratio
Preparation	Generate clinical data	129 rows x 3 columns
	Generate JSON conversational dialogue	129 unique dialogue user_id
	Train-Validation-Test Split	70-15-15



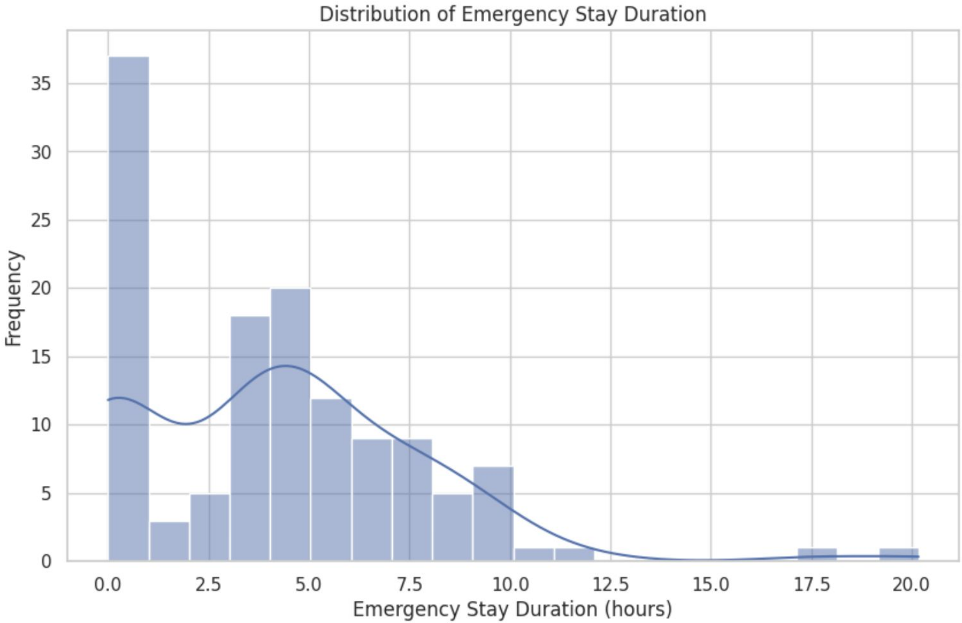
04

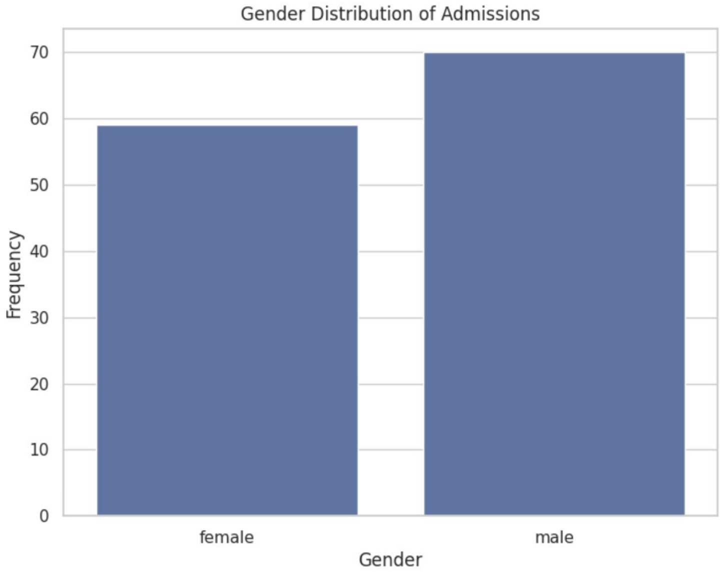
Data Analytics Results



The graph displays the distribution of admission types, showing a significantly higher frequency of emergency admissions compared to elective and urgent admissions.

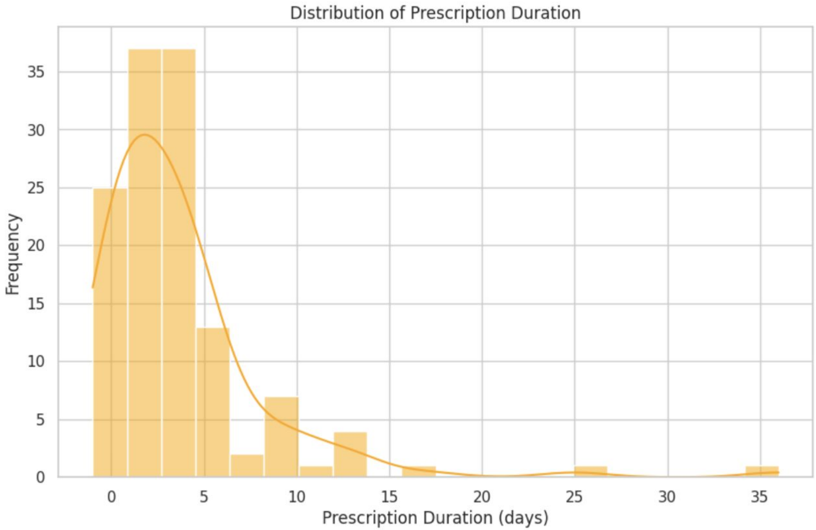
The graph illustrates the distribution of emergency stay durations, showing a high frequency of shorter stays (up to around 2.5 hours), with frequencies decreasing as the duration increases, and the majority of stays lasting less than 10 hours.

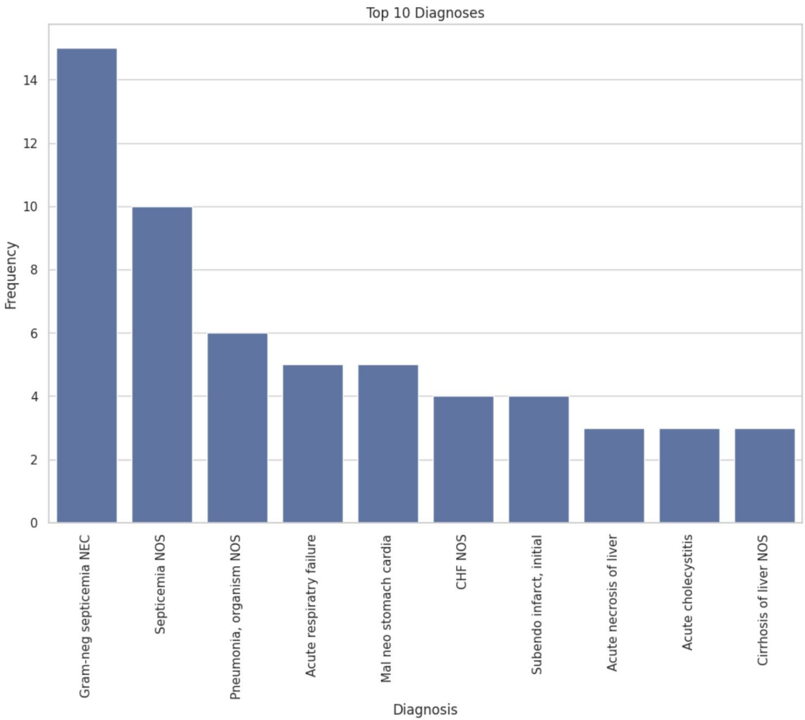




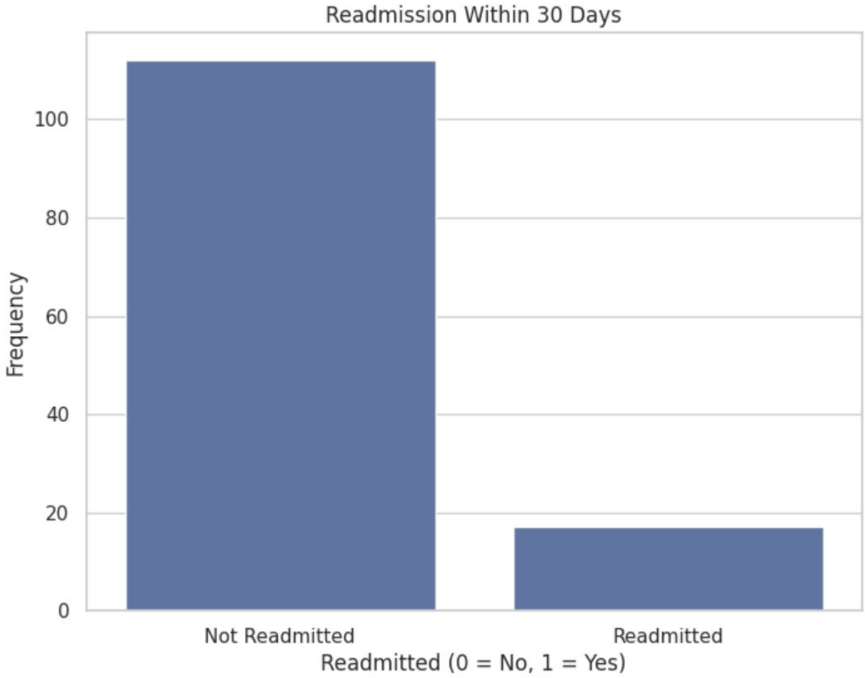
The graph depicts the gender distribution of admissions, showing a slightly higher frequency of male admissions compared to female admissions.

The graph shows the distribution of prescription durations, with the majority of prescriptions lasting between 0 and 5 days, and a sharp decline in frequency as the duration increases beyond 10 days.

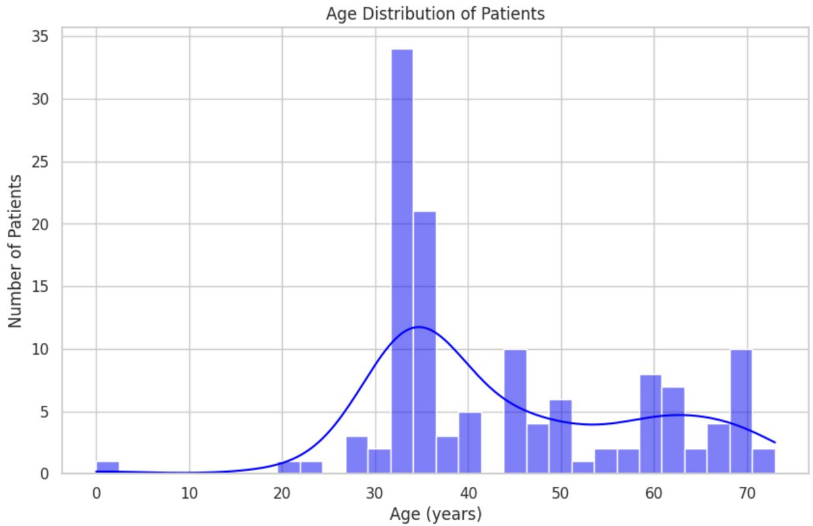




Top 10 diagnoses, with "Gram-neg septicemia NEC" being the most frequent diagnosis among the listed conditions.

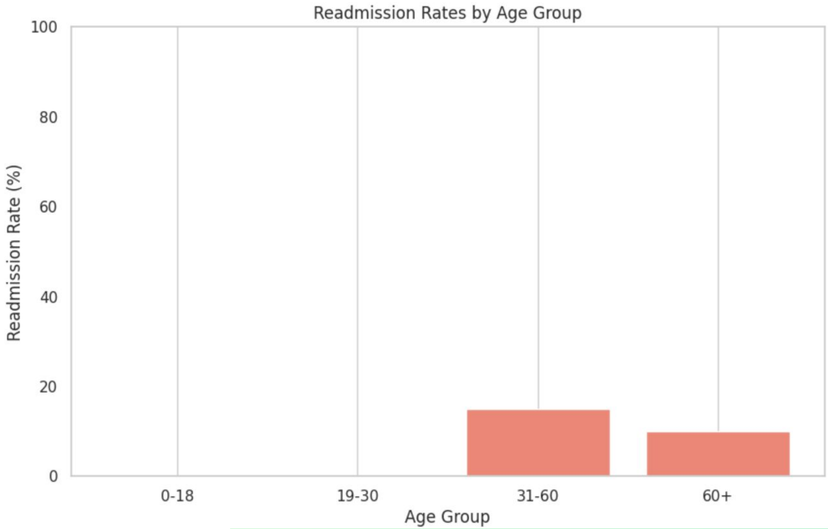


Patients who were not readmitted within 30 days and those who were, with far fewer readmissions. - before class balancing



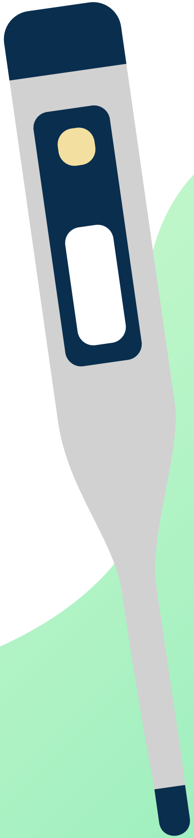
The graph displays the age distribution of patients, showing a notable peak around the age of 30 years, with the distribution declining on either side, indicating fewer patients in both the younger and older age groups.

The graph presents readmission rates by age group, indicating that the readmission rates for the age groups 19-30 and 60+ are higher compared to the other age groups, which show significantly lower or no data.



05

Model Development



LLM Models Proposed

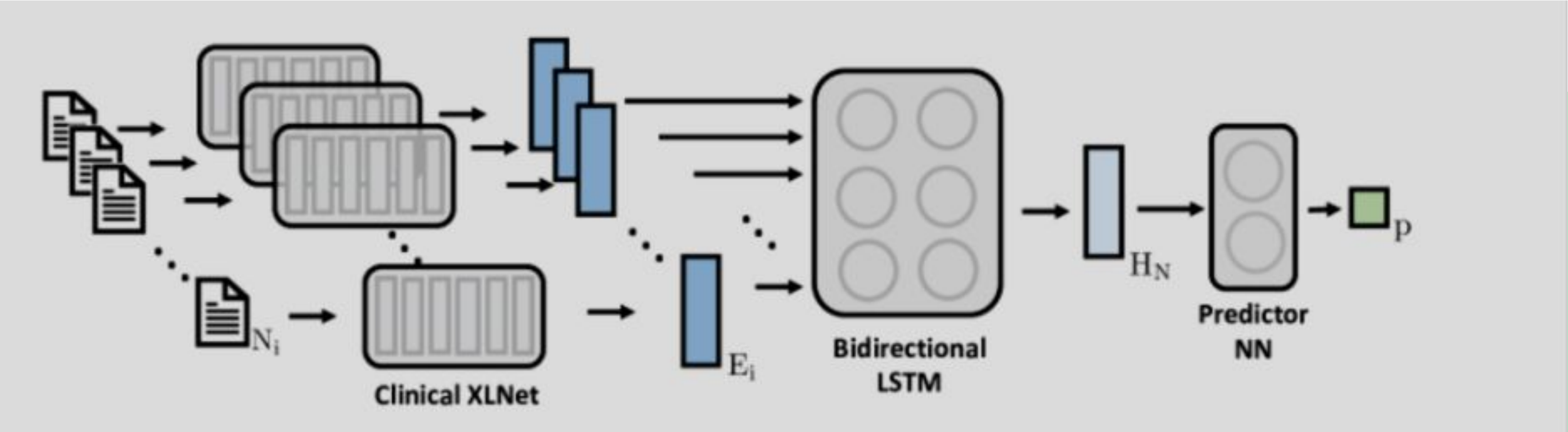
1. **ClinicalBert** - Specifically fine-tuned on clinical text, making it well-suited for understanding medical language and patient narratives relevant to readmission risks.
2. **Clinical XLNet** - Effectively captures contextual dependencies in clinical notes and enables for bidirectional context, increasing patient data interpretation and interaction.
3. **GPT-4o** - Provides significant natural language interpretation and generating skills, allowing it to analyse patient communications and generate predictive insights into readmission risks.
4. **Clinical BigBird** is a transformer model optimized for long documents such as clinical texts, thus efficiently handling electronic health records and medical literature.
5. **PubMedBERT** is a domain-specific BERT model that has been pre-trained on biomedical and clinical texts from PubMed and was designed to improve the NLP tasks in the biomedical domain.

Clinical XLNet

Model Selection: Chose XLNet for its strength in handling long text, ideal for detailed patient summaries in readmission prediction.

Key Improvements: Hyperparameter tuning and stratified cross-validation improved accuracy and balanced performance across classes.

Innovations: Created a custom dataset structure and used AUC-ROC and F1-score metrics for detailed performance insights.

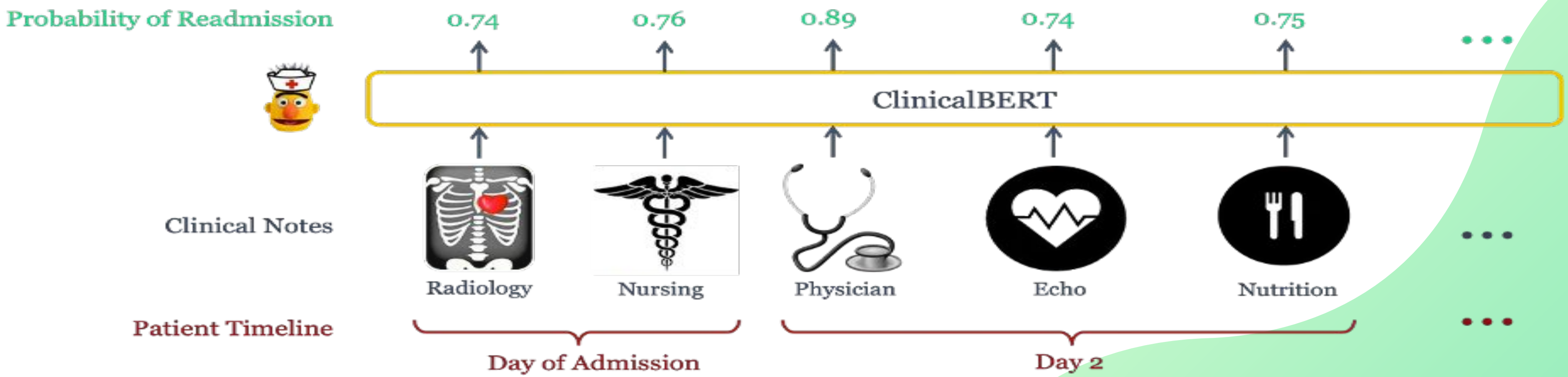


ClinicalBert

Model Selection: Selected ClinicalBERT for its medical text pre-training, enhancing readmission risk prediction from patient summaries.

Key Improvements: Fine-tuned for readmission prediction, optimized hyperparameters, and used dynamic padding for efficient data handling.

Innovations: Added AUC-ROC, F1-score, and confusion matrix visualizations to improve performance insights, particularly for imbalanced classes.

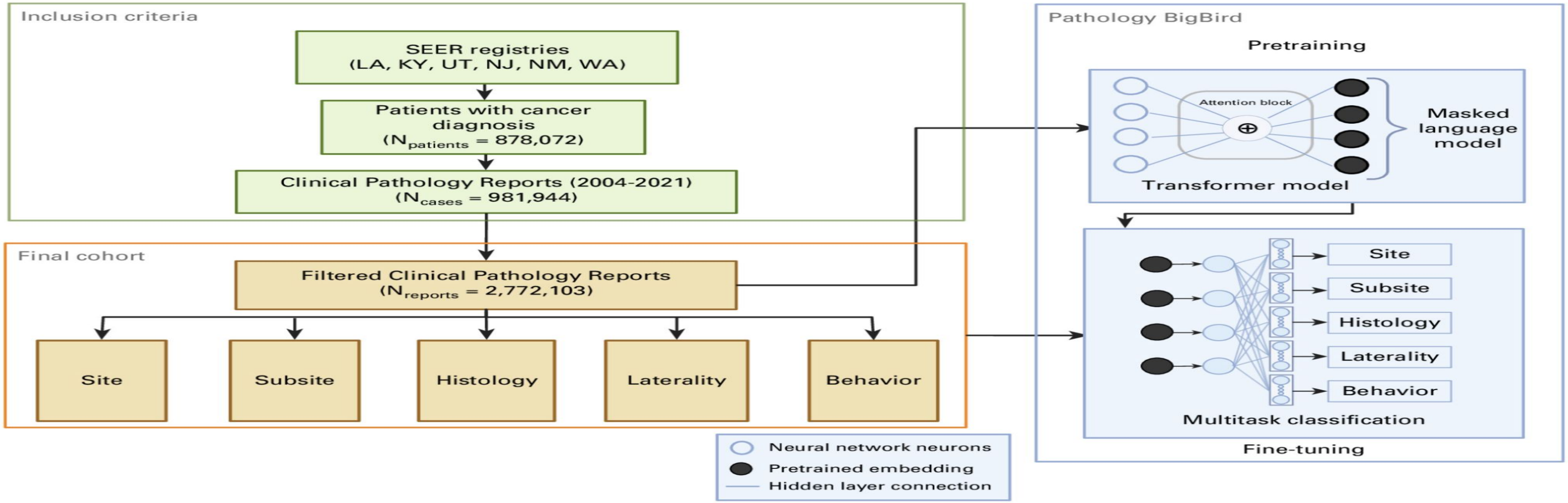


Clinical BigBird

Model Selection: Chose Clinical Bigbird for handling long, unstructured clinical data efficiently with sparse attention.

Key Improvements: Custom preprocessing for medical terminology, hyperparameter tuning, and domain-specific vocabulary training enhanced long-text performance.

Innovations: Developed an evaluation pipeline using AUC-ROC and precision-recall curves for detailed performance assessment.

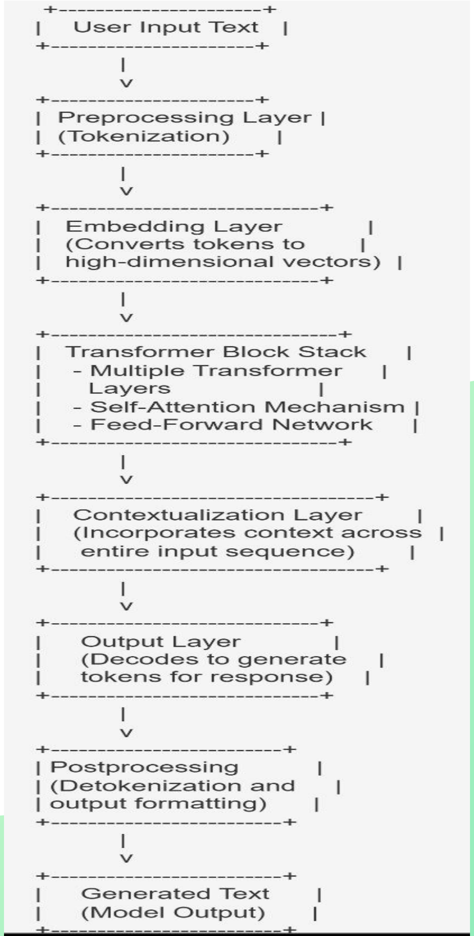


GPT-4 Turbo

Model Selection: Chose GPT-4 Turbo for its interpretive depth in medical case analysis, enhancing predictive accuracy for patient readmissions within an ensemble of KNN and logistic regression.

Key Improvements: Optimized caching reduced latency for recurring queries; hyperparameter tuning and adaptive ensemble weighting improved balanced performance, validated via cross-validation.

Innovations: Developed a hybrid ensemble of KNN, logistic regression, and GPT-4 Turbo, weighted for optimal readmission prediction. Evaluated using F1-score and recall to improve both accuracy and interpretability.

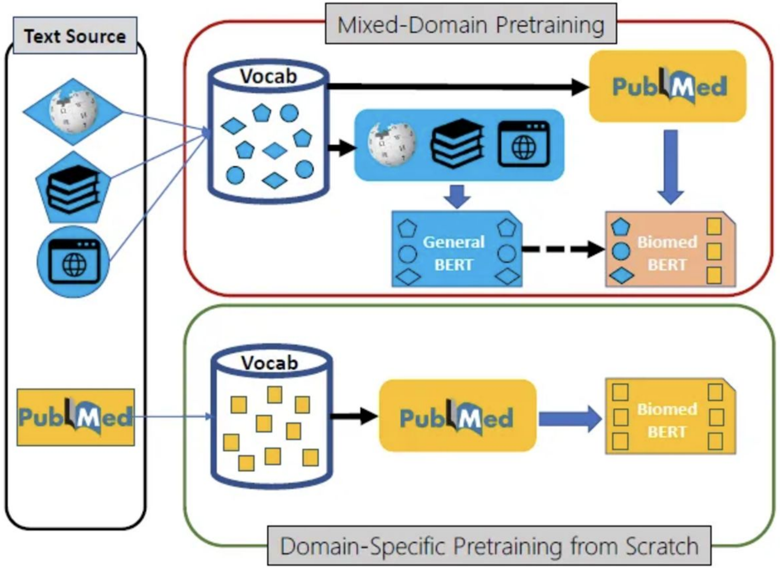


Pub MedBert

Model Selection: PubMedBERT was chosen for its domain-specific pre-training on biomedical literature, enabling precise analysis of clinical text and EHRs for readmission prediction.

Key Improvements: Fine-tuned on healthcare datasets, PubMedBERT enhanced pattern recognition for readmission risks, validated with F1-scores and precision-recall metrics.

Innovations: Integrated PubMedBERT into an ensemble framework, enabling real-time data analysis, personalized care planning, and reduced hospital readmission rates.



05

Model Evaluation Results and Comparisons

Evaluation Metrics

- **Precision:** Proportion of correct positive predictions.
- **Recall:** Model's ability to identify true positives.
- **F1-Score:** Balance of precision and recall, useful for class imbalance.
- **Accuracy:** Overall correctness of predictions.
- **GLEU Score:** measures the similarity between a generated text and a reference text

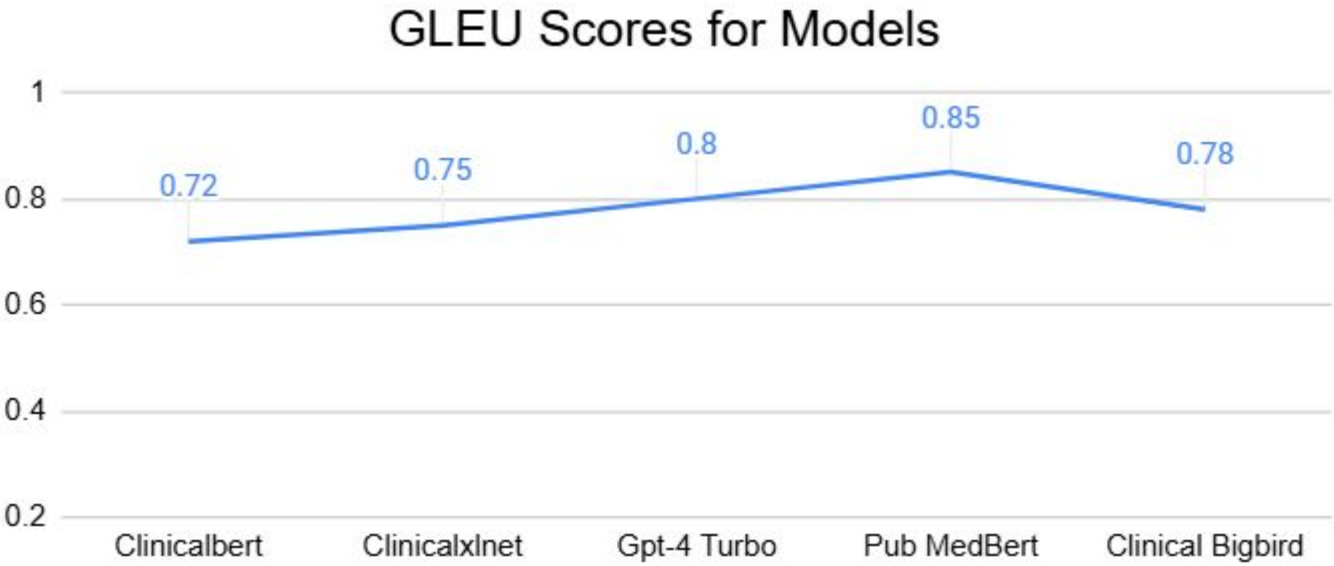
Pub MedBert Model Results

Note: The above graph shows training and validation loss of the Pub MedBert Model

Model Evaluation Results

Models	Precision	Recall	F1	Accuracy
ClinicalBERT	0.94	0.97	0.96	0.96
ClinicalXLNet	0.9658	0.9649	0.9648	0.9649
GPT-4 Turbo	0.92	0.9	0.91	0.91
ClinicalBigBird	0.9563	0.9551	0.955	0.9551
PubMedBert	0.9779	0.9852	0.9816	0.9795

Model Evaluation Results



Pub Medbert Model Results

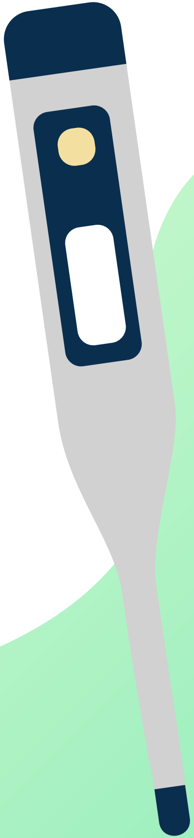
Prediction on Patient Clinical Summary

Clinical Summary: I am a male with the diagnosis of shortness of breath admitted as a emergency case. Discharged to HOME HEALTH CARE, covered by Private insurance. I received 18 prescriptions, stayed 2.1026 days in the ICU as a metavision patient, medication was administered via IV as MAIN and the drugs included sodium chloride 0.9% flush. My emergency room stay lasted 459.0 days.', 'model': 'Hello, I am here to help assess your readmission risk. Please provide your gender, diagnosis, admission type, discharge location, insurance, number of prescriptions, ICU duration, ICU type, route of medication administration, drug type, specific drugs taken, and emergency stay duration.

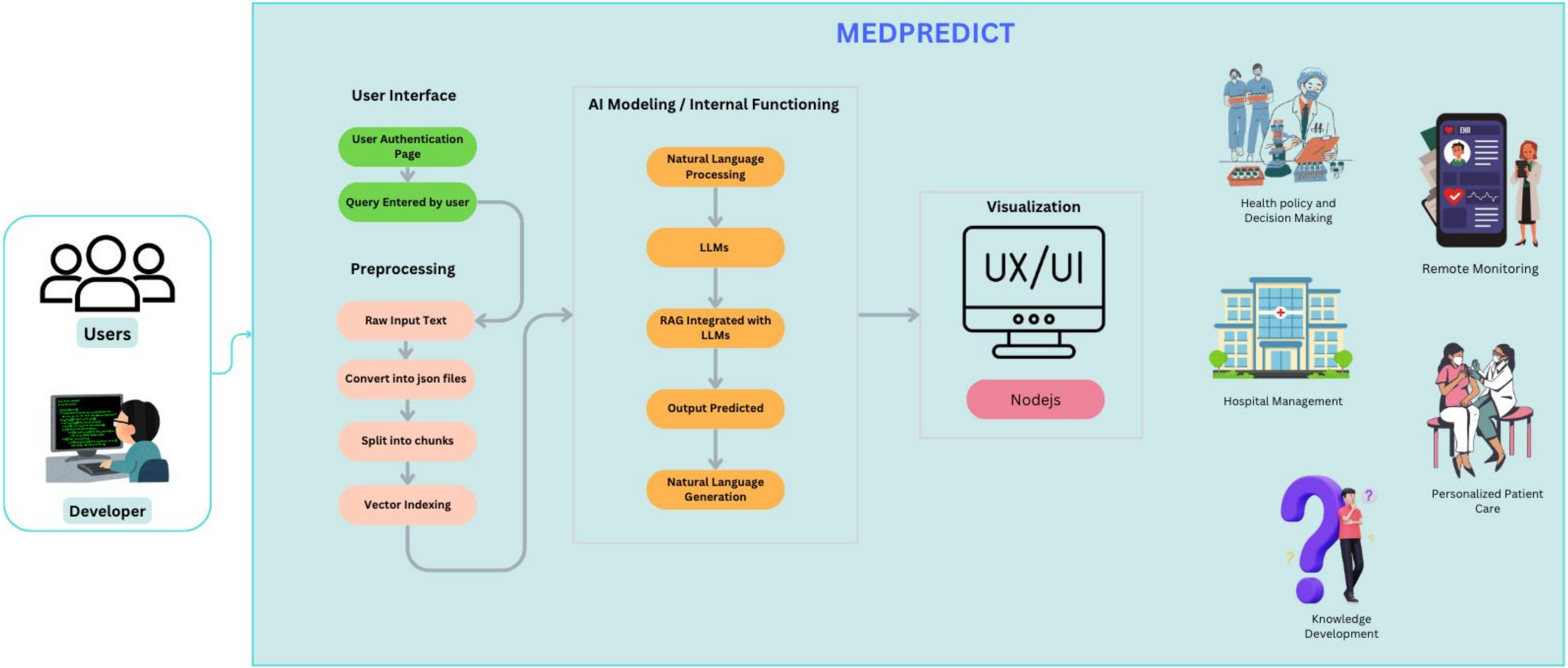
```
Prediction: High risk of readmission
Dialogue: {'patient': 'I am a male with the diagnosis of shortness of breath admitted as a emergency case.
```

06

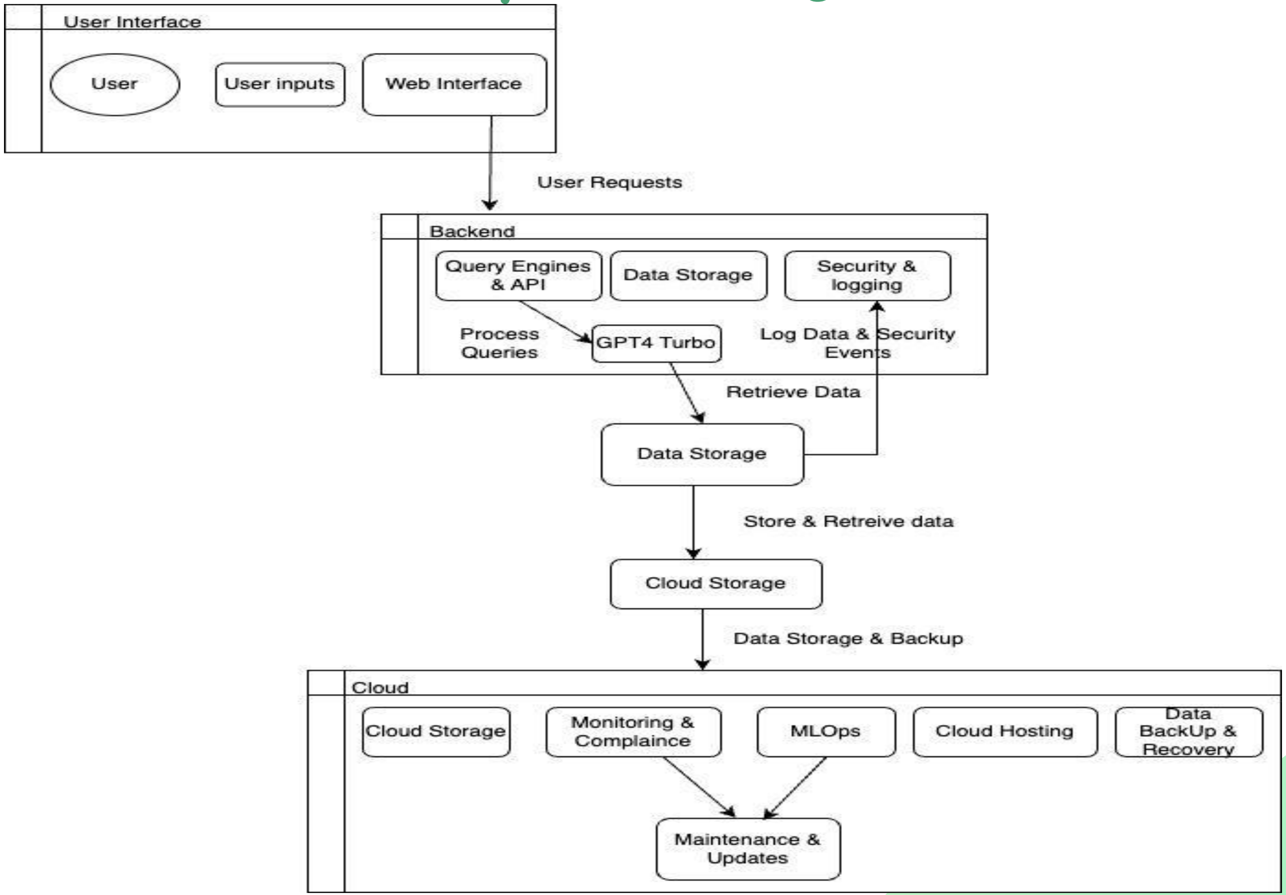
System Development



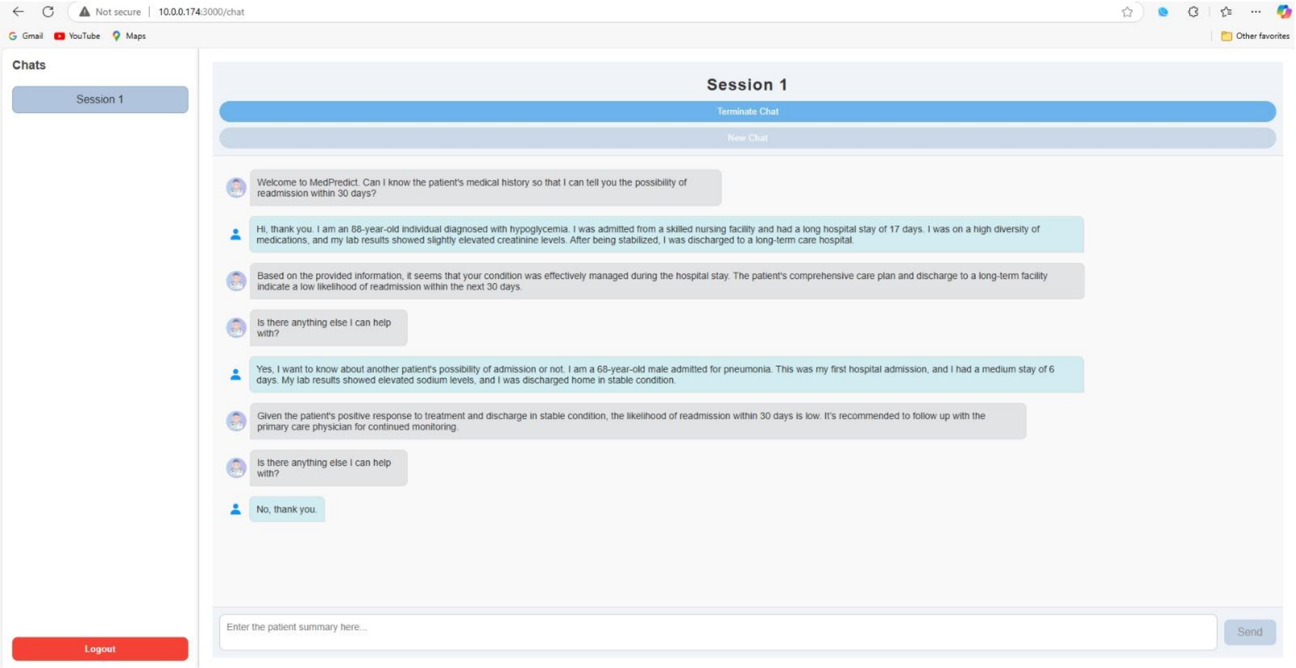
Overall System Design



UI System Design



User Interface



07 DEMO

Demo Link- https://drive.google.com/drive/folders/1pBQJCbELGlUqxk5tDTmBQhnPFwl-P4Zh?usp=drive_link

08 References

A. Pathak, Om Shree, M. Agarwal, Shek Diya Sarkar, and Anupam Tiwary, “Performance Analysis of LoRA Finetuning Llama-2,” Dec. 2023, doi: https://doi.org/10.1109/iementech60402.2023.10423400

B. Saha, S. Lisboa, and S. Ghosh, “Understanding patient complaint characteristics using contextual clinical BERT embeddings,” *arXiv (Cornell University)*, Jul. 2020, doi: https://doi.org/10.1109/embc44109.2020.9175577.

C. Baechle, A. Agarwal, Ravi Behara, and X. Zhu, “A cost sensitive approach to predicting 30-day hospital readmission in COPD patients,” Jan. 2017, doi: https://doi.org/10.1109/bhi.2017.7897269.

G. Althari and M. Alsulmi, “Exploring Transformer-Based Learning for Negation Detection in Biomedical Texts,” *IEEE Access*, vol. 10, pp. 83813–83825, 2022, doi: https://doi.org/10.1109/access.2022.3197772.

H. Ali, J. Qadir, T. Alam, Mowafa Househ, and Z. Shah, “ChatGPT and Large Language Models in Healthcare: Opportunities and Risks,” Sep. 2023, doi: https://doi.org/10.1109/aibthings58340.2023.10291020.

THANK YOU