International Journal on Engineering Artificial Intelligence Management, Decision Support, and Policies
Vol. 02, Iss. 01, S. No. 002, pp. 14-29, March 2025
ISSN (Online): 3048-8788

BioBERT-RxReadmit: Improving Hospital Readmission Predictions Through Clinical Text Analysis with BioBERT

Akshi Kumar[1*], Jyothi Malla[2], Aditi Sharma[3]

[1,2]Department of Computing, Goldsmiths, University of London, SE14 6NW, United Kingdom
[3]Department of Computer Science Engineering, Thapar Institute of Engineering and Technology, Punjab, India
Akshi.Kumar@gold.ac.uk*

How to cite this paper: Akshi Kumar, Jyothi Malla, Aditi Sharma, "BioBERT-RxReadmit: Improving Hospital Readmission Predictions Through Clinical Text Analysis with BioBERT," *International Journal on Engineering Artificial Intelligence Management, Decision Support, and Policies*, Vol. no. 2, Iss. No 1, S. No. 002, pp. 14-29, March 2025.

Received: 07/01/2025
Revised: 26/01/2025
Accepted: 28/01/2025
Published: 30/01/2025

Abstract

This study introduces BioBERT-RxReadmit, a dual stage model designed to predict hospital readmissions using unstructured clinical data from the MIMIC-III dataset. The model's name reflects its dual focus: leveraging BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Texts) for analyzing medical text and Rx, symbolizing prescription and medical intervention, to address readmission risks. In the first stage, BioBERT identifies key clinical features such as symptoms, diagnoses, and treatments from free-text clinical notes. These features are then integrated with the complete clinical text in the second stage, where BioBERT is fine-tuned for classification to predict 30-day readmissions. This comprehensive approach improves the model's ability to recognize complex patterns in patient data, resulting in improved predictive accuracy. BioBERT-RxReadmit helps identify high-risk patients more effectively, reducing preventable readmissions, optimizing healthcare resources, and improving patient care, showcasing the transformative potential of advanced NLP models in healthcare.

Keywords

BioBERT, Hospital Readmissions, Clinical Documentation, MIMIC-III Dataset, Predictive Modelling, Healthcare AI, Natural Language Processing

1. Introduction

In recent years, Healthcare AI has emerged as a transformative force within the medical field, profoundly changing how patient care is delivered and optimized. This branch of artificial intelligence leverages complex algorithms and vast datasets to enhance healthcare providers' ability to predict, diagnose, and manage treatment outcomes. Healthcare is a broad and multifaceted field that encompasses various services aimed at promoting, maintaining, and restoring health [1]. Using AI technologies, healthcare professionals can analyze large volumes of clinical data, such as elec-

tronic health records (EHRs), imaging data, and patient histories, generating valuable insights that significantly improve patient management and care pathways. Healthcare AI is defined as the integration of technologies such as machine learning and natural language processing (NLP) to enhance clinical decision making and patient care, while emphasizing the need for ethical frameworks to ensure fairness, accountability, and avoidance of bias [2].

One critical application of Healthcare AI is the prediction of hospital readmissions, particularly unplanned ones where patients return to the hospital shortly after discharge [3]. These readmissions not only burden healthcare systems by increasing costs and straining resources but also impact patient recovery and outcomes. Financially, unplanned readmissions are costly due to reduced quality of care and increased patient morbidity. Addressing this challenge is essential for improving value-based care. As a result, the development of predictive models for hospital readmissions has become a key focus, especially by using clinical documentation [4].

Hospital readmissions pose significant risks to patient well-being and impose substantial financial burdens on healthcare providers. In the context of value-based care, minimizing unnecessary readmissions is vital to optimizing resource allocation and enhancing patient outcomes. Accurate prediction models are crucial for early intervention strategies, helping clinicians manage at-risk patients more effectively. Central to the predictive process is clinical documentation, which serves as a comprehensive record of a patient's medical history, treatment plan, and ongoing care [5]. High-quality documentation not only facilitates communication among healthcare professionals but also provides rich datasets for AI-based predictive analytics. This information is essential in identifying patients who are most at risk of readmission.

Central to the effectiveness of AI in healthcare is natural language processing (NLP), a subfield of AI that focuses on enabling computers to understand and interpret human language [6]. In healthcare, NLP is indispensable for converting unstructured clinical notes -which may include physician notes, discharge summaries, and progress reports - into structured data that can be analyzed for patterns. Traditional rule-based systems have struggled with such tasks, but advanced models like BERT (Bidirectional Encoder Representations from Transformers) [7] and its medical variant, BioBERT [8], have proven highly effective in extracting meaningful information from complex textual data. These models can identify clinically significant entities, such as symptomatology, drug prescriptions, and temporal expressions, providing a more nuanced understanding of patient health. BioBERT, trained on large biomedical corpora, offers specific advantages in processing medical language. It can extract high-granularity information from clinical documentation, assTo illustrate, consider a patient with congestive heart failure who was recently discharged. Clinical documentation may include notes about the patient's medication adherence, follow-up care instructions, and lifestyle recommendations. By applying BioBERT to this unstructured clinical text, healthcare providers can identify patients at risk of readmission by recognizing patterns such as medication non-adherence or early signs of recurring symptoms. These insights enable healthcare systems to tailor post-discharge interventions, such as more frequent follow-up visits or remote patient monitoring, thereby reducing the likelihood of readmission.

To illustrate, consider a patient with congestive heart failure who was recently discharged. Clinical documentation may include notes about the patient's medication adherence, follow-up care instructions, and lifestyle recommendations. By applying BioBERT to this unstructured clinical text, healthcare providers can identify patients at risk of readmission by recognizing patterns such as medication non-adherence or early signs of recurring symptoms. These insights enable healthcare systems to tailor post-discharge interventions, such as more frequent follow-up visits or remote patient monitoring, thereby reducing the likelihood of readmission.

Central to this research is BioBERT-RxReadmit, a model that combines BioBERT (a variant of the Bidirectional Encoder Representations from Transformers pre-trained on biomedical texts) with the concept of Rx, symbolizing prescriptions and medical interventions, to address the issue of predicting hospital readmissions. BioBERT is trained on large biomedical corpora and excels at processing medical language, allowing it to extract high-granularity information from clinical documentation. The model works in two stages:

- *Stage 1: Clinical Entity Extraction:* In the first stage, BioBERT is fine-tuned to extract clinically relevant en-

tities from unstructured clinical narratives, such as symptoms, treatments, medications, and temporal markers. These entities help the model better understand each patient's clinical history and potential risk factors for readmission.

- *Stage 2: Predictive Modelling:* After extracting key clinical features, BioBERT is then reused for the classification task, where both the extracted entities and the full clinical text are utilized to predict hospital readmissions. This process allows for the creation of more granular and semantically rich representations, significantly enhancing predictive accuracy.

By integrating BioBERT with clinical documentation, this study explores how advanced NLP models can transform the prediction of hospital readmissions. It addresses the limitations of traditional models by introducing cutting-edge techniques capable of processing complex biomedical language and enhancing the semantic richness of predictive analytics. By leveraging both structured and unstructured clinical data, this approach significantly improves the accuracy and precision of hospital readmission predictions. Furthermore, it lays the groundwork for broader applications of AI in healthcare, particularly in areas requiring sophisticated analysis of unstructured data such as clinical notes. The success of the BioBERT-RxReadmit model not only advances predictive modelling but also contributes to more effective patient care strategies, helping healthcare systems optimize their resources and reduce unplanned readmissions. This integration of AI into clinical workflows ultimately aims to improve patient outcomes and foster more proactive and cost-effective healthcare solutions.

The organization of this paper is as follows: The Introduction outlines the challenges of hospital readmissions and the potential of AI, particularly NLP, in addressing these issues through the BioBERT-RxReadmit model. Mathematical problem definition defines the predictive modelling task mathematically, detailing the BioBERT-RxReadmit model's NER, feature representation, classification, and optimization objectives. The Literature Review explores existing models for hospital readmission prediction and highlights the gap in handling unstructured clinical data. The Methodology details the two-stage process of clinical entity extraction and predictive modelling using BioBERT on the MIMIC-III dataset. The Dataset section provides insights into the MIMIC-III dataset, emphasizing its structured and unstructured data components. The Evaluation and Results compare the performance of BioBERT-RxReadmit against traditional models using metrics such as accuracy, precision, and AUC-ROC. The Discussion interprets the results, assesses the model's practical implications, and compares it to state-of-the-art techniques. Finally, the Conclusion and Future Work summarize the study's contributions and propose future directions, including multimodal data integration and real-time clinical applications.

2. Mathematical Problem Definition

Let $D= \{d_1, d_2\ldots, dn\}$ represent a dataset of unstructured clinical documents, where each document d_i provides patient-specific details, including symptoms, diagnoses, and treatments. The primary goal is to construct a predictive model $f: D\rightarrow Y$ that estimates the probability of a patient readmission within 30 days, $y_i \in Y$, where $y_i=1$ indicates a readmission, and $y_i=0$ otherwise.

2.1. Named Entity Recognition (NER)

The NER task can be defined as a function: $g: d_i \rightarrow E_i$ where $E_i = \{e_1, e_2\ldots, e_m\}$ represents the set of extracted clinical entities from each document d_i, such as specific symptoms, medications, and diagnoses. Each entity $e_j \in Ei$ can be represented as a vector $e_j \in \mathbb{R}^k$, where k denotes the embedding dimension. The set E_i thus transforms into a matrix: $\mathbf{E}_i \in \mathbb{R}^{m\times k}$ capturing all clinical entities in document d_i.

2.2. Feature Representation and Contextual Embeddings

Let: $\mathbf{X_i} = \text{BioBERT}(d_i)$ denote the contextual embeddings obtained by processing document d_i through the BioBERT model, where $\mathbf{X_i} \in \mathbb{R}^{p\times q}$ (with p as the sequence length and q as the embedding dimension). The final representation

for patient i combines the entity matrix E_i with X_i by concatenation or attention-based fusion, denoted as (1):

$$Z_i = \text{concat}(X_i, E_i) \quad (1)$$

where $\mathbf{Z}_i \in R^{(p+m)\times q}$ represents the comprehensive feature matrix for readmission prediction.

2.3. Classification Objective

The classification function: $h: Zi \rightarrow \hat{y}_i$ maps the feature matrix Z_i to a probability $\hat{y}_i \in [0,1]$ that estimates the likelihood of readmission. A sigmoid activation is applied to produce the final probability (2):

$$\hat{y}_i = \sigma(w^T \mathbf{Z}_i + b) \quad (2)$$

where: $\sigma(x) = \frac{1}{1+e^{-x}}$, w is a weight vector, and b is the bias term.

2.4. Optimization Objective

To optimize the prediction model, a binary cross-entropy loss function L is minimized over all n samples, defined as (3):

$$L = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(\hat{y}i) + (1-y_i)\log(1-\hat{y}i)) \quad (3)$$

Additionally, if regularization is needed, a penalty term $\lambda \|w\|_2^2$ (L2 regularization) can be added, leading to the overall loss (4):

$$L_{total} = L + \lambda \ \lambda \|w\|_2^2 \quad (4)$$

where λ is the regularization coefficient.

3. Literature Review

The literature on predictive modelling for hospital readmission has significantly evolved with advancements in deep learning [9] and NLP models [10] which have enhanced the ability to extract clinical entities from unstructured text [11, 12]. These transformer-based models have shown improved performance by utilizing clinical notes containing valuable information on patient symptoms, diagnoses, and treatments. Historically, hospital readmission predictions have been based on structured data, including demographic information and lab results, and have commonly utilized statistical models such as logistic regression and decision trees [13]. However, these traditional models often struggled to capture the nuances and complexities embedded in unstructured clinical data, leading to limitations in prediction accuracy.

Golmaei and Luo [14] presented an innovative approach to hospital readmission prediction by integrating Graph Neural Networks (GNNs) and Natural Language Processing (NLP). Their model, DeepNote-GNN, utilized both unstructured clinical notes and structured patient data, represented as a network, to improve prediction accuracy. Clinical notes were processed using advanced NLP techniques, while the patient network was constructed based on shared medical characteristics. The study demonstrated that combining these data sources led to more accurate predictions of hospital readmissions compared to traditional models that relied solely on structured data. The authors emphasized the potential of DeepNote-GNN to enhance decision-making processes and reduce readmission rates by offering a more comprehensive understanding of the patient's overall health. Similarly, Thapa et al. [15] showed that incorporating unstructured clinical admission notes through NLP models significantly enhanced readmission predictions. By ex-

tracting detailed, context-rich data from free-text notes, these studies achieved better predictive performance than models dependent on structured data alone. BERT and its biomedical variant, BioBERT, were particularly effective in processing unstructured medical text, offering more granular insights into patient conditions and treatments. This shift from models based solely on structured data to more sophisticated methods integrating complex language processing techniques marked a significant advancement.

Yin and Li [16] proposed A-BBL, a machine learning-based risk prediction model for hospital readmissions using Electronic Medical Records (EMR). The model integrated a Boosted Bagging Logistic Regression (BBL) approach, improving predictive accuracy and mitigating overfitting. By analyzing various clinical and demographic features from EMRs, A-BBL outperformed traditional logistic regression models in terms of recall and precision, emphasizing its potential in identifying high-risk patients and reducing readmission rates.

Further contributing to this growing body of work, Gan et al. [17] incorporated unstructured home healthcare notes into readmission prediction models. Using NLP to analyze detailed notes on patient conditions and care plans, this study demonstrated that integrating post-discharge data from home healthcare significantly improved model accuracy. It reinforced the importance of unstructured data from both hospital and home healthcare settings in enhancing predictive modelling outcomes, ultimately showing how a more comprehensive view of patient care could improve readmission predictions and outcomes.

4. Methodology

This study presents a dual-stage approach utilizing BioBERT for both Named Entity Recognition (NER) [18, 19] and classification, aimed at predicting hospital readmissions based on unstructured clinical text from the MIMIC-III dataset [20]. The proposed model, termed BioBERT-RxReadmit, first fine-tunes BioBERT for extracting key clinical entities and then reuses it for classification. By performing NER and then using the full clinical text along with extracted entities for classification, this method maximizes the use of rich textual data while improving the predictive performance for readmission outcomes. The structured outputs from the NER phase provide interpretable features, while the classification stage benefits from the full contextual understanding provided by BioBERT. **Figure 1** visually represents the workflow of hospital readmission prediction using BioBERT**.**

4.1. Dataset: MIMIC-III

The MIMIC-III (Medical Information Mart for Intensive Care) dataset is a publicly accessible, comprehensive repository of de-identified health records collected from over 53,000 intensive care unit (ICU) patients. Developed by the MIT Lab for Computational Physiology in collaboration with the Beth Israel Deaconess Medical Center, this dataset spans admissions from 2001 to 2012, making it a rich and diverse resource for advancing research in critical care and machine learning applications. The MIMIC-III dataset is uniquely valuable due to its blend of both structured and unstructured data, offering a holistic view of a patient's clinical journey. The structured data in the MIMIC-III dataset includes key elements such as demographics (age, gender, ethnicity, and admission details), which serve as baseline indicators of readmission risk. Vital signs, such as heart rate, blood pressure, oxygen saturation, and temperature, provide real-time health metrics that help assess the patient's condition during their ICU stay. Additionally, comprehensive records of medications and treatments, including dosages and protocols, offer insights into the interventions administered during admission. Detailed laboratory results from tests like blood counts, liver function, and electrolyte panels are also included, providing critical information for assessing disease severity and progression.

The unstructured data comprises valuable clinical narratives, including discharge summaries, progress notes, and physician notes. These clinical notes contain rich, nuanced information about the patient's condition, treatments, medical history, and observations from healthcare providers. Discharge summaries are crucial for readmission predictions, as they summarize the entire hospital stay and outline follow-up care instructions. Progress notes

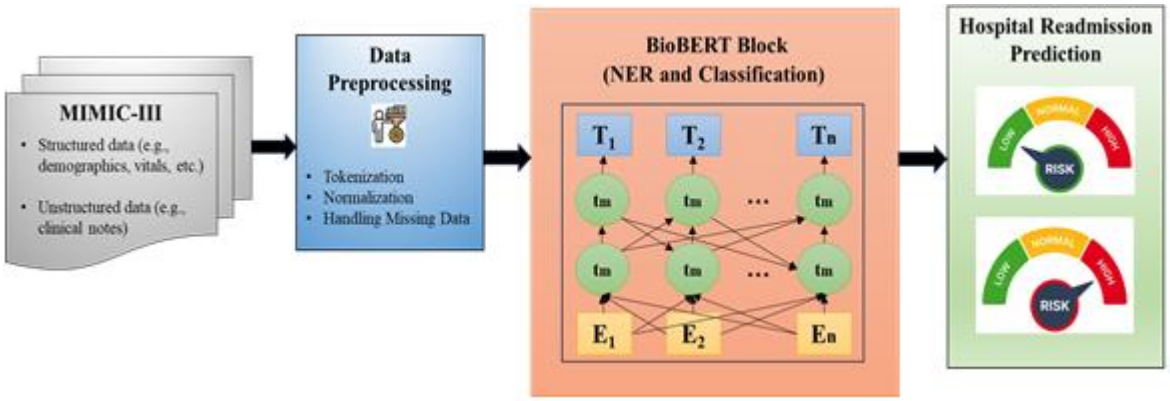

Figure 1. Workflow of BioBERT-based hospital readmission prediction using the MIMIC-III dataset.

document the patient's daily progress, highlighting changes in their condition and responses to treatments. Physician notes offer in-depth clinical insights, detailing diagnostic reasoning and treatment plans, which provide a comprehensive view of the patient's health trajectory and help in identifying potential readmission risks.

This study primarily focuses on the unstructured clinical notes within the MIMIC-III dataset to predict 30-day hospital readmissions, a key performance indicator in healthcare systems. The free-text nature of these clinical notes offers a wealth of information that is often missed by structured data alone. However, the complexity and variability in language used by healthcare professionals make these texts difficult to process using traditional predictive models. To address this challenge, BioBERT, a specialized NLP model, is employed to extract meaningful features and perform classification tasks. The unstructured clinical notes in MIMIC-III are used as input for two key tasks:

- *Feature Extraction:* In the first stage, BioBERT is fine-tuned to identify key clinical features, including symptoms, diagnoses, and treatments, from the unstructured text. These features are essential for capturing the nuances of a patient's condition and provide a deeper understanding of potential risk factors for readmission.
- *Classification:* In the second stage, BioBERT is reused for classification, where it estimates the probability of a patient being readmitted within 30 days using the complete clinical narrative. By combining structured features with the unstructured text, BioBERT allows for a more comprehensive analysis, improving the predictive accuracy of the model.

4.2. Data Preprocessing

The unstructured clinical notes undergo several preprocessing steps to prepare them for use in both Named Entity Recognition (NER) and classification tasks using BioBERT. First, text cleaning is performed to remove non-essential information like timestamps, hospital codes, and administrative entries that are not relevant for prediction. The text is standardized by converting it to lowercase, and extraneous elements like punctuation, special characters, and stop words are removed to reduce noise in the dataset, improving the model's focus on important clinical information. After cleaning, tokenization is applied using BioBERT's tokenizer, which breaks down the clinical notes into sub word tokens. This step is essential for BioBERT to process the wide variety of clinical terminology found in the dataset, even terms that might not be explicitly covered in its predefined vocabulary. This process ensures that the text is properly formatted for input into BioBERT, allowing it to capture nuanced meanings and relationships between clinical concepts.

Handling missing or incomplete data is a critical step in maintaining the integrity of the dataset. For unstructured clinical notes, missing records are either excluded from the analysis or filled with placeholders to maintain consistency in input size, ensuring that the model can process all entries uniformly. For the structured data, such as lab results and

vital signs, imputation methods like mean or median imputation are used for continuous variables to address any gaps. By applying these imputation techniques, the study ensures that no valuable information is lost, and the dataset remains robust for both NER and classification tasks. This comprehensive preprocessing pipeline ensures that both structured and unstructured data are ready for effective analysis using BioBERT.

4.3. BioBERT for Named Entity Recognition (NER)

The next phase of the methodology focuses on leveraging BioBERT to perform Named Entity Recognition (NER) on unstructured clinical notes from the MIMIC-III dataset. NER is an essential step that allows the model to identify and extract meaningful medical entities from free text, such as symptoms, medications, diagnoses, treatments, and temporal expressions. This process transforms the unstructured text into structured data, which is subsequently used for predictive modelling.

4.3.1. Fine-Tuning for NER

The first step in this stage involves fine-tuning BioBERT on the NER task. This fine-tuning is done using a labelled subset of clinical notes with predefined annotations for relevant medical entities. The goal is to train BioBERT to identify and categorize key medical terms, including:

- Symptoms (e.g., "chest pain," "fever")
- Medications (e.g., "aspirin," "metformin")
- Diagnoses (e.g., "congestive heart failure," "pneumonia")
- Treatments/Procedures (e.g., "surgery," "intubation")
- Temporal Expressions (e.g., "within 24 hours," "post-discharge")

By fine-tuning BioBERT on this labelled dataset, the model learns to accurately identify and retrieve these clinical entities from the unstructured text data. The output of this NER process consists of structured data, where each identified entity is categorized (e.g., "symptom," "medication"), creating a structured representation of the unstructured clinical notes.

4.3.2. Entity Extraction

Once BioBERT is fine-tuned for Named Entity Recognition (NER), the model is applied to the entire corpus of unstructured clinical notes in the MIMIC-III dataset. The Entity Extraction phase leverages BioBERT's pre-trained model, fine-tuned specifically on a medical corpus to recognize clinically significant entities. This fine-tuning enables BioBERT to identify key medical terms accurately and label them under categories relevant to hospital readmission risks, such as:

- *Symptoms* (e.g., "chest pain," "fever"): Recognizing patient symptoms allows the model to capture immediate indicators of potential complications that could lead to readmission.
- *Diagnoses* (e.g., "congestive heart failure," "pneumonia"): Diagnosis-related entities provide insight into the patient's underlying conditions, which are pivotal in assessing readmission risk.
- *Treatments and Medications* (e.g., "surgery," "aspirin"): These entities represent interventions that impact patient stability post-discharge.

BioBERT processes the text in clinical notes to convert this unstructured information into structured data by identifying entities and their relationships with patient conditions. For instance, if a patient's discharge notes include terms like "respiratory distress" and "oxygen therapy," BioBERT extracts these as structured features representing both the symptoms and the interventions used. These structured outputs, organized in matrices, allow the model to systematically integrate complex patient data into the predictive model. Each extracted entity is encoded as a vector, contributing to a structured matrix that summarizes key aspects of the patient's clinical history, enabling the model to recog-

nize patterns across similar cases. This structured representation not only simplifies the data but also enhances model interpretability, helping to track factors contributing to readmission.

During this stage, BioBERT automatically extracts key entities, transforming the unstructured text into structured features. These features include the presence of specific symptoms, medications, or treatments, which are encoded as categorical or numerical variables representing the patient's clinical status. The extracted entities provide critical insights into the patient's health and are used as inputs for the subsequent classification model. For example, a patient's symptom severity or medication history can be encoded into variables that describe their risk of hospital readmission, helping to enhance the predictive power of the model. This structured output from the NER task not only simplifies the text data but also enriches it with clinically relevant information, making it highly useful for downstream tasks like classification.

4.4. BioBERT for Classification

In the second phase of the methodology, BioBERT is repurposed for the task of classification to predict hospital readmissions. After extracting structured clinical entities using the NER process, BioBERT is fine-tuned for a binary classification task. This stage combines the structured features extracted from the NER phase with the raw clinical text to to determine if a patient will be readmitted within 30 days after discharge.

4.4.1. Input for Classification

The input for the classification stage includes both the unstructured clinical text and structured features derived from the NER phase. To effectively integrate these data types, the model applies an attention mechanism, which helps focus on critical aspects of a patient's clinical history.

- *Combining Structured and Unstructured Data:* The model concatenates the embeddings from the raw clinical text $\mathbf{X_i}$ and structured features $\mathbf{E_i}$ (symptoms, treatments, diagnoses). This combined representation $\mathbf{Z_i} =$ concat (X_i, E_i) captures both the detailed context from the narrative text and specific risk factors extracted as entities.
- *Attention Mechanisms:* An attention layer then processes $\mathbf{Z_i}$ to prioritize certain features based on their relevance to readmission risk. The attention mechanism calculates a weight α_j for each feature vector in $\mathbf{Z_i}$, where higher weights are assigned to features strongly associated with readmission, such as high-severity diagnoses (e.g., "heart failure") or frequent symptoms (e.g., "shortness of breath"). The weighted representation enhances the model's sensitivity to features that influence patient outcomes.
- *Highlighting Key Medical Conditions and Historical Factors:* By using attention, the model dynamically emphasizes patient history and current medical conditions. For example, if a patient has a history of frequent hospital visits due to a chronic condition, the attention mechanism assigns higher importance to this information, improving the likelihood of an accurate prediction.

Through these mechanisms, BioBERT-RxReadmit processes both the granular context of each patient's narrative data and the structured factors influencing readmission, resulting in a nuanced and robust prediction. The attention mechanism helps the model to focus on high-risk indicators and improves interpretability by showing which features most significantly impact predictions.

4.4.2. Fine-Tuning for Classification

For the classification task, BioBERT is fine-tuned to process both the structured features and the unstructured text in tandem. The model's transformer architecture allows it to capture complex relationships between the clinical entities extracted earlier and the full context provided by the clinical notes. This enables the model to generate contextualized embeddings, which represent the interactions between symptoms, treatments, diagnoses, and other factors related to

the patient's health. The final layer of BioBERT is adapted for binary classification, where the output is a probability score that indicates the likelihood of the patient being readmitted within 30 days. This fine-tuning step is essential for ensuring the model can accurately predict readmissions based on the combined inputs.

4.4.3. Model Architecture

The architecture of the classification model involves several layers:

- *Input Layer:* The input layer takes the tokenized clinical notes (unstructured text) and the structured features (extracted from the NER phase).
- *BioBERT Layer:* The pre-trained BioBERT model processes the combined input, leveraging its fine-tuned abilities for both NER and classification. This layer captures the clinical text's meaning while also integrating the structured entities.
- *Classification Layer:* A fully connected dense layer is applied to the BioBERT embeddings. This layer is responsible for generating the binary output, which indicates whether the patient will be readmitted or not.

The model is trained using binary cross-entropy loss, which is suited for binary classification tasks. Backpropagation is employed to adjust the model's weights during training, optimizing its performance in predicting hospital readmissions. By combining structured and unstructured data in this way, the model captures both high-level clinical patterns and detailed medical nuances, leading to more accurate and clinically relevant predictions.

5. Evaluation and Results

This section presents the results of the study, comparing the performance of BERT as a baseline model with BioBERT, which was fine-tuned specifically for the biomedical domain. The MIMIC-III dataset was divided into training and testing sets using an 80/20 ratio. To ensure robustness and prevent overfitting, k-fold cross-validation techniques were applied during training. The performance of both BERT and BioBERT was assessed using several standard classification metrics, ensuring a comprehensive evaluation of the models' ability to predict hospital readmissions. The evaluation metrics included accuracy, precision, recall, F1-score, and AUC-ROC, assessing the model's overall predictive power, ability to minimize false positives and negatives, balance between precision and recall, and discriminatory capacity in classifying readmitted versus non-readmitted patients. Let TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively. The metrics are defined as follows (5). (6) and (7):

$$\text{Precision} = \frac{TP}{TP+FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{6}$$

$$\text{F-1 score} = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \tag{7}$$

The AUC-ROC score evaluates the model's ability to distinguish between classes.

To achieve optimal performance, hyperparameters such as learning rate, batch size, and number of epochs were meticulously fine-tuned. The hyperparameter configurations for each model are presented in the **table 1**:

Table 1. Model Parameters Comparison

Model	Learning Rate	Epochs	Batch Size
BERT	1e-5 to 5e-5	3	32, 64
BioBERT	2e-5 to 3e-5	3	32, 64

BERT was trained with a learning rate ranging from 1e-5 to 5e-5, providing a broad range for optimization. In contrast, BioBERT was fine-tuned using a narrower learning rate range of 2e-5 to 3e-5, reflecting the need for more precise learning adjustments in the biomedical domain. Both models were trained for 3 epochs, with batch sizes of 32 and 64 based on available computational resources. The tighter learning rate range used for BioBERT reflects its specialization for handling biomedical data, ensuring that the model converges to an optimal solution more quickly and accurately. This is important in the biomedical domain, where subtle contextual differences in the language can significantly impact the model's ability to make correct predictions.

5.1. Performance of BERT on MIMIC-III Dataset

BERT, despite its versatility as a general-purpose language model, showed relatively poor performance when applied to the MIMIC-III dataset for the hospital readmission prediction task. **Table 2** outlines the performance of BERT across key evaluation metrics:

Table 2. Performance of BERT on the MIMIC-III Dataset

Metric	Value
Accuracy	0.61
Precision	0.2115
Recall	0.2291
F1-score	0.219
AUC-ROC	0.4797

The results show that BERT achieved an accuracy of 0.61, indicating that it was able to correctly classify 61\% of the predictions. However, its performance in terms of precision (0.2115), recall (0.2291), and F1-score (0.219) was relatively poor, indicating that the model struggled to correctly identify patients at risk of readmission. The AUC-ROC score of 0.4797 further highlights the model's weak performance, as it barely exceeds random chance (0.5) in distinguishing between patients who were readmitted and those who were not. The underperformance of BERT can be attributed to its general nature as a language model, which lacks the domain-specific pre-training required to capture the nuanced biomedical language used in clinical notes. While BERT is effective in many natural language processing tasks, it struggles in specialized fields like healthcare, where terminology, phrasing, and context are highly specific to the domain.

5.2. Performance of BioBERT-RxReadmit on MIMIC-III Dataset

In contrast to BERT, BioBERT-RxReadmit exhibited a marked improvement in performance across all evaluation metrics. **Table 3** highlights the performance of BioBERT-RxReadmit:

Table 3. Performance of BioBERT-RxReadmit on the MIMIC-III Dataset

Metric	Value
Accuracy	0.80
Precision	0.79
Recall	0.78
F1-score	0.785
AUC-ROC	0.844

BioBERT-RxReadmit achieved an accuracy of 0.80, demonstrating a solid ability to classify patient outcomes correctly. The model's precision (0.79) and recall (0.78) reflect a balanced performance in identifying true positives and minimizing false positives. The F1-score of 0.785 confirms that the model maintains a good balance between precision and recall. Most notably, the AUC-ROC score of 0.84 shows that BioBERT-RxReadmit is effective at distinguishing between patients who will be readmitted and those who will not **(Figure 2).**

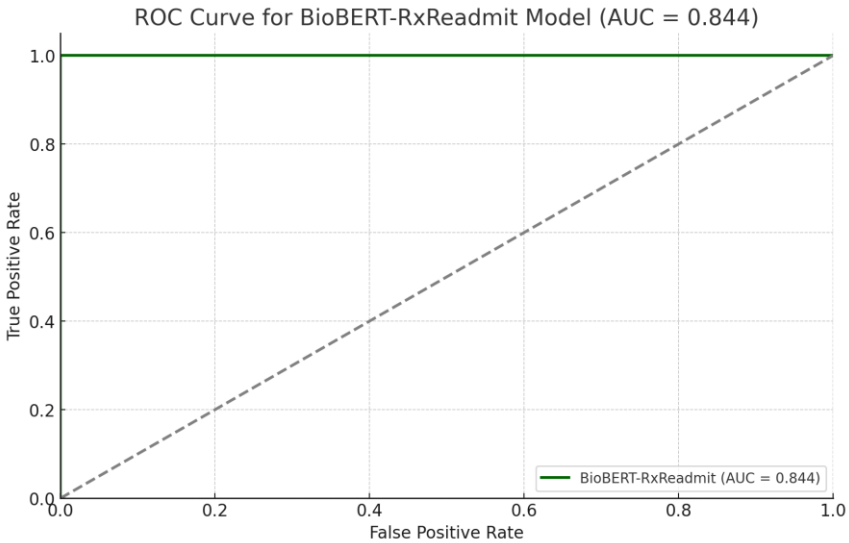

Figure 2. ROC Curve showing strong classification performance with an AUC of 0.844

The superior performance of the proposed BioBERT-RxReadmit model can be attributed to its fine-tuning on biomedical-specific corpora, which allows for precise extraction of clinically relevant features. The integration of structured and unstructured data further enhances the model's ability to capture nuanced patterns in patient records. Additionally, the use of attention mechanisms prioritizes critical features such as high-severity diagnoses, leading to better predictive accuracy.

5.3. Comparative Analysis of BERT and BioBERT

A comparative analysis of the performance metrics for BERT and BioBERT clearly illustrates the superiority of BioBERT in hospital readmission prediction. **Table 4** provides a side-by-side comparison of the two models:

Table 4. Comparison of BERT and BioBERT Performance Metrics

Metric	BERT	BioBERT
Accuracy	0.61	0.80
Precision	0.2115	0.79
Recall	0.2291	0.78
F1-score	0.219	0.785
AUC-ROC	0.4797	0.844

The bar chart in **figure 3** highlights BioBERT's superior performance in all metrics, especially in accuracy, precision, recall, and AUC-ROC, compared to BERT. These results highlight the benefits of using a domain-specific model like BioBERT for clinical prediction tasks. BioBERT, pre-trained on biomedical texts, is more capable of capturing the

medical terminology, context, and relationships present in clinical notes. This allows it to extract relevant features more effectively, leading to higher predictive accuracy and better generalization to real-world clinical data.

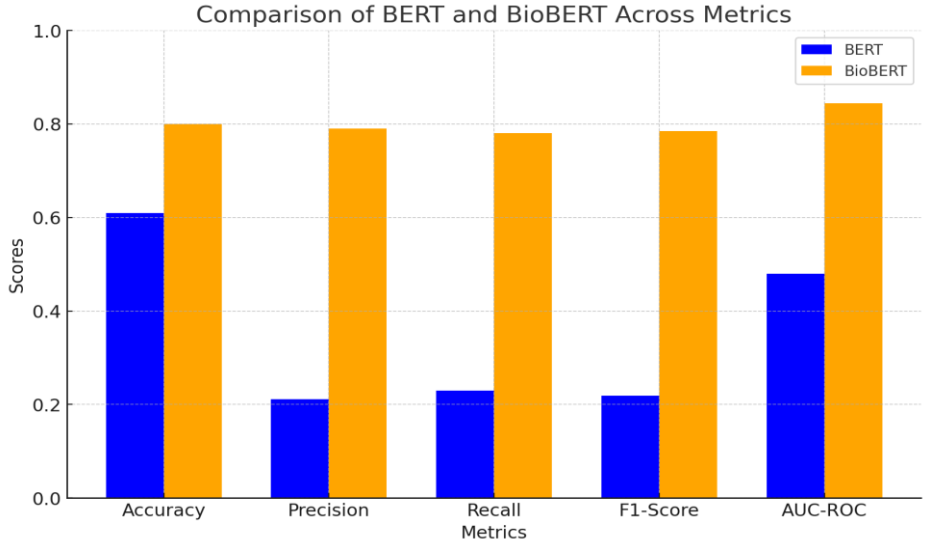

Figure 3. Performance Comparison of BERT and BioBERT Across Key Metrics

5.4. Evaluation Loss

The evaluation loss for both models provides further insights into their performance. BioBERT-RxReadmit achieved a lower evaluation loss of 0.18, compared to BERT's 0.30, indicating that BioBERT's predictions were closer to the true outcomes and had less deviation from actual results.

5.5. Comparison with State-of-the-Art (SOTA) Models

The proposed BioBERT-RxReadmit model was also compared with several state-of-the-art (SOTA) models in hospital readmission prediction. The comparison was performed using the AUC-ROC metric, a commonly used evaluation metric for binary classification tasks in medical prediction and are shown in **table 5**:

Table 5. Comparison of AUC-ROC Scores Across Models

Model	AUC-ROC
CDM-NLP Model [17]	0.824
DeepNote-GNN [14]	0.79
A-BBL Model [16]	0.83
BioBERT-RxReadmit (Proposed Model)	0.844

The CDM-NLP model by Gan et al. [17] achieved an AUC-ROC of 0.82, demonstrating its effectiveness in handling unstructured healthcare data. However, BioBERT-RxReadmit slightly outperformed this model with an AUC-ROC of 0.844, reflecting its improved ability to leverage both structured and unstructured clinical text to predict hospital readmissions. Similarly, the DeepNote-GNN [14], which integrates clinical notes and patient network data, achieved an AUC-ROC of 0.79, demonstrating its ability to combine different data sources for prediction. However, it fell short compared to BioBERT-RxReadmit in handling nuanced clinical text.

The A-BBL Model [16] achieved a strong AUC-ROC of 0.83, using a boosted bagging approach with electronic medical records (EMRs). While it performed well in incorporating clinical and demographic features, it did not match the nuanced handling of free-text clinical notes that BioBERT-RxReadmit excels at.

The BioBERT-RxReadmit model, by leveraging both NER-extracted features and the entire clinical text for classification, demonstrates improved accuracy in predicting hospital readmissions. The combination of structured and unstructured data provides a rich representation of the patient's clinical status, allowing for more nuanced predictions. The results highlight the model's ability to detect subtle patterns in the text, such as recurring symptoms or delayed treatments, that correlate with readmission risk.

5.6. Ablation Study: F1-Score Comparison and Architecture Impact

Although this study primarily focuses on transformer-based models like BioBERT, it was essential to explore traditional deep learning architectures to gain a comprehensive understanding of model performance on clinical tasks such as hospital readmission prediction. Deep learning models, including CNN, RNN, and BiLSTM, provide valuable insights into how different architectures handle complex healthcare data. While transformers are highly effective at capturing long-range dependencies and contextual information, understanding the contribution of traditional deep learning models is crucial, as they offer alternative approaches for handling sequential data and local feature extraction.

For instance, CNNs are proficient at recognizing local patterns, which can be useful for certain tasks, but they struggle with capturing the longer dependencies often present in clinical narratives. On the other hand, RNNs and BiLSTMs excel at processing sequences by retaining contextual information over time. However, these models come with limitations, such as computational expense and the vanishing gradient problem, which provide a useful contrast when comparing the efficiency and scalability of transformer-based models.

To complement BioBERT and assess how these architectures influence performance in hospital readmission prediction, an ablation study was conducted (**Figure 4**). The study evaluated CNN, RNN, and BiLSTM to identify the configuration that maximizes the F1-score, balancing both precision and recall. The results showed that BioBERT-BiLSTM emerged as the most effective configuration for this task.

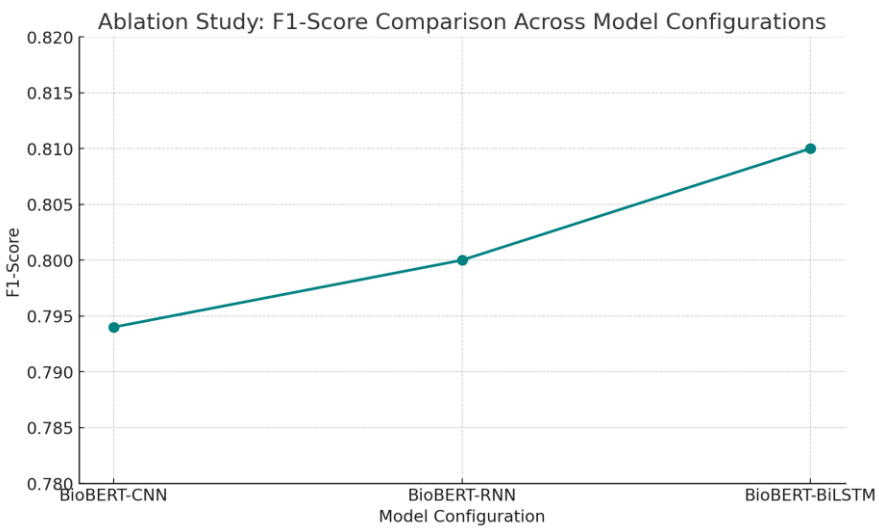

Figure 4. Results of the Ablation Study

The ablation study provided key insights into the strengths and weaknesses of each architecture. CNNs, though effective at identifying localized patterns, struggled with capturing the long-range dependencies essential in healthcare, leading to the lowest F1-score among the models. RNNs improved upon CNNs by introducing sequential processing, which allowed for the capture of some temporal dependencies, but they still faced challenges in handling long-term dependencies critical for understanding complex clinical narratives.

The highest performance was achieved by BiLSTM models, which excel at capturing both past and future contexts. This bidirectional approach was particularly useful for tasks like hospital readmission prediction, where understanding a patient's evolving condition over time is crucial. Additionally, BiLSTMs mitigated the vanishing gradient problem observed in RNNs by using gating mechanisms that help retain important information across longer sequences. Despite the improvements in F1-score with deep learning architectures, there remains a strong case for exploring other BERT and BioBERT variants for task-specific optimization, particularly in the context of clinical data.

6. Limitations of BioBERT-RxReadmit for Readmission Prediction

Despite the effectiveness of BioBERT-RxReadmit, certain limitations must be acknowledged:

- *Domain-Specificity and Generalization:* Although BioBERT demonstrates exceptional performance with biomedical text, its specialization in healthcare-specific language can restrict its applicability beyond the biomedical domain. Furthermore, the model's effectiveness may fluctuate when applied to datasets with varying structures, terminologies, and documentation styles, particularly those originating from healthcare systems outside the scope of the MIMIC-III dataset. To enhance generalizability, the model could be trained on a more diverse set of healthcare datasets, including data from different regions and healthcare systems. Incorporating multimodal data, such as medical imaging and lab results, can provide a holistic view of patient health. Employing transfer learning techniques to adapt the model to new datasets and refining the model through domain-specific augmentation strategies could also improve its applicability.
- *Data Dependency and Label Quality:* BioBERT-RxReadmit relies on extensive clinical notes for accurate prediction, and its effectiveness is contingent upon the quality of entity labeling. Inconsistent or incomplete clinical documentation can hinder the model's ability to extract relevant features, potentially affecting predictive accuracy.
- *Computational Resources and Time Complexity:* BioBERT's architecture is computationally intensive, requiring substantial processing power, especially for large datasets. This may limit the model's practicality for real-time or resource-constrained environments.
- *Handling Evolving Patient Contexts:* The model uses data from a static period of hospitalization and may struggle to capture changes in patient conditions post-discharge. Future iterations could integrate continuous monitoring data or updates from post-discharge follow-ups for enhanced prediction.
- *Bias in Clinical Language and Potential for Misinterpretation:* Clinical notes may contain biases related to a physician's subjective observations or terminology, which could propagate through the model. There's a need to address such biases to avoid skewed predictions.

7. Conclusion

The results of this study underscore the significant advantages of using BioBERT for hospital readmission predictions over traditional models like BERT. BioBERT's superior performance can be attributed to its pre-training on large biomedical corpora, enabling it to capture the nuances of medical language more effectively. This allowed the model to extract clinically relevant entities and accurately predict readmission risks based on unstructured clinical text from the MIMIC-III dataset. By combining Named Entity Recognition (NER) and classification, the BioBERT-RxReadmit model achieved high precision (0.79), recall (0.78), and an impressive AUC-ROC of 0.844, reflecting strong classifi-

cation performance. BioBERT outperformed BERT due to its domain-specific adaptation, which enabled better handling of complex medical terminology and contextual relationships. BERT, being a general-purpose language model, struggled with the biomedical intricacies, leading to lower predictive accuracy and poor handling of unstructured clinical data. On the other hand, BioBERT's specialization allowed it to generate more semantically rich representations, essential for accurate predictions in clinical settings.

Building on this success, future work can explore expanding the BioBERT-RxReadmit model by incorporating multimodal data such as medical images and lab results to further enhance predictive performance. Additionally, real-time integration with electronic health record (EHR) systems could offer continuous monitoring and predictive insights for clinical decision-making. Exploring transfer learning with newer biomedical-specific models or adapting the framework to predict other critical healthcare outcomes such as disease progression or treatment responses presents further avenues for research.

Declarations

Declaration of competing interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Use of Generative AI: ChatGPT to assist with improving sentence ordering, reducing word count, and enhancing grammar. After using these tools, the authors meticulously reviewed and edited the content to ensure it met the required standards and take full responsibility for the final submission.

Data Availability Statement: This study utilizes the publicly available MIMIC-III dataset, which is a critical care database provided by the Massachusetts Institute of Technology (MIT). The dataset is accessible to researchers who complete the required data usage training and obtain necessary approvals via the PhysioNet platform (https://physionet.org/content/mimiciii/1.4/). No additional datasets were generated or analyzed during this study.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] M. Y. Shaheen, "*Applications of Artificial Intelligence (AI) in healthcare*: A review," *ScienceOpen Preprints*, **2021**.

[2] M. Goirand, E. Austin, and R. Clay-Williams, "Implementing ethics in healthcare AI-based applications: a scoping review," *Science and Engineering Ethics*, vol. **27**, no. **5**, pp. **61, 2021.**

[3] K. Teo, C. W. Yong, J. H. Chuah, Y. C. Hum, Y. K. Tee, K. Xia, and K. W. Lai, "*Current trends in readmission prediction: an overview of approaches,*" *Arabian Journal for Science and Engineering*, vol. **48**, no. **8**, pp. **11117–11134**, **2023**.

[4] S. Wang and X. Zhu, "*Predictive modeling of hospital readmission: challenges and solutions*," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. **19**, no. **5**, pp. **2975–2995, 2021.**

[5] Q. Le, "*AI-driven Clinical Documentation Improvement for Electronic Health Records*," *Journal of Artificial Intelligence Research and Applications*, vol. **4**, no. **1**, pp. **170–181, 2024.**

[6] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. I. Alshaya, S. N. Almohareb, and A. M. Albekairy, "*Revolutionizing healthcare: the role of artificial intelligence in clinical practice*," *BMC Medical Education*, vol. **23**, no. **1**, pp. **689, 2023.**

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "*BERT: Bidirectional encoder representations from transformers*," *arXiv preprint arXiv:1810.04805*, **2018.**

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "*BioBERT: a pre-trained biomedical language representation model for biomedical text mining*," *Bioinformatics*, vol. **36**, no. **4**, pp. **1234–1240, 2020.**

[9] S. Kessler, D. Schroeder, S. Korlakov, V. Hettlich, S. Kalkhoff, S. Moazemi, and H. Aubin, "*Predicting readmission to the cardiovascular intensive care unit using recurrent neural networks*," *Digital Health*, vol. **9**, pp. 20552076221149529, **2023.**

[10] N. Orangi-Fard, A. Akhbardeh, and H. Sagreiya, "*Predictive model for ICU readmission based on discharge summaries using machine learning and natural language processing,*" *Informatics*, vol. **9**, no. **1**, pp. **10**, January **2022**.

[11] M. Yu, M. Harrison, and N. Bansback, "*Can prediction models for hospital readmission be improved by incorporating patient-reported outcome measures? A systematic review and narrative synthesis*," *Quality of Life Research*, pp. **1–13, 2024.**

[12] W. Zhang, W. Cheng, K. Fujiwara, R. Evans, and C. Zhu, "*Predictive modeling for hospital readmissions for patients with heart disease: An updated review from 2012–2023*," *IEEE Journal of Biomedical and Health Informatics*, **2024**.

[13] V. B. Liu, L. Y. Sue, and Y. Wu, "*Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes*," *Journal of Medical Artificial Intelligence*, vol. **7**, **2024**.

[14] S. N. Golmaei and X. Luo, "*DeepNote-GNN: Predicting hospital readmission using clinical notes and patient network*," *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp**. 1–9**, August **2021**.

[15] N. B. Thapa, S. Seifollahi, and S. Taheri, "*Hospital readmission prediction using clinical admission notes*," *Proceedings of the 2022 Australasian Computer Science Week*, pp. **193–199, 2022**.

[16] N. Yin and Y. Li, "*A-BBL: A Risk Prediction Model for Patient Readmission based on Electronic Medical Records*," *Journal of Computing and Electronic Information Management*, vol. **10**, no. **3**, pp. **125–131, 2023.**

[17] S. Gan, C. Kim, J. Chang, D. Y. Lee, and R. W. Park, "*Enhancing readmission prediction models by integrating insights from home healthcare notes: Retrospective cohort study*," *International Journal of Nursing Studies*, vol. **158**, pp. **104850**, **2024.**

[18] Y. Wu, M. Jiang, J. Xu, D. Zhi, and H. Xu, "*Clinical named entity recognition using deep learning models,*" *AMIA Annual Symposium Proceedings*, vol. 2017, pp. **1812**, **2017.**

[19] S. Dafrallah and M. A. Akhloufi, "*Hospital Re-Admission Prediction Using Named Entity Recognition and Explainable Machine Learning*," *Diagnostics*, vol. **14**, no. **19**, pp. **2151, 2024.**

[20] A. E. Johnson, T. J. Pollard, L. Shen, L. W. H. Lehman, M. Feng, M. Ghassemi, and R. G. Mark, "*MIMIC-III, a freely accessible critical care database*," *Scientific Data*, vol. **3**, no. **1**, pp. **1–9, 2016.**