ClinicalML: Using Machine Learning for clinical outcome prediction

Shaswat Patel
New York University
spp9399@nyu.edu

Vishvesh Trivedi
New York University
vt2369@nyu.edu

Jeet Patel
New York University
jmp10073@nyu.edu

Team 15

Abstract

Improving patient care is highly dependent on accurate clinical outcome predictions, especially for tasks such as mortality and length of stay. Electronic Health Records (EHR) contain essential patient information that can be harnessed by machine learning models to predict these outcomes. However, EHR data can be noisy and without effective feature engineering, models risk underfitting. Current frameworks utilize large language models that offer little explainability to clinicians. To address this dual problem, we propose ClinicalML, a lightweight pipeline designed to de-noise EHR data, extract relevant features, and integrate Machine Learning(ML) models for predicting critical patient outcomes. Our pipeline achieved a 0.58 macro-F1 score for clinical mortality and a 0.33 macro-F1 score for predicting length of stay. The pipeline also offers explainability to further develop trust between clinicians and ML models.

1 Introduction

Hospitals produce massive structured and unstructured clinical data. Intensive Care Units (ICU) are increasingly using data and prediction models to support their decision making. Early prediction of patient trajectory and clinical outcomes can prompt early treatments, improve outcomes, and even save patient lives in the ICU.[1, 2] Clinicians assess patients' condition in the intensive care unit using a broad range of information modalities such as vitals, patient history, laboratory results, etc. Clinical outcome predictions, such as mortality and length of stay, are among the most commonly modeled outcomes in the ICU. However, most studies focus on the use of vital signs. These systems often fail when vital signs for a patient are unavailable, which is common when patients have just been admitted to the ICU.[3] In this work, we focus on using only the textual data from the EHR available at the time of admission to predict two clinical outcomes: Mortality Prediction (MP) and Length Of Stay (LoS).

BERT-based[4] pre-trained models have demonstrated an impressive ability to extract task-related information from noisy, unstructured data. However, BERT models have common pitfalls: (i) they are data intensive and require a large amount of supervised data, particularly in the context of medical outcomes, to achieve reliable metrics [5]; and (ii) their black-box nature limits interpretability and transparency[6]. Machine Learning (ML) methods, on the other hand, offer reliable, stable, and explainable models that can achieve comparable results to BERT models when effective feature engineering techniques are employed. Thus, we base our work on classical ML algorithms and pose the following questions:

1. *Can traditional Machine Learning models achieve comparable results when compared to BERT based state-of-the-art approaches from admission notes?*
2. *Can traditional Machine Learning models identify and interpret common risk factors? And what are the common limitations of ML models?*

To address the aforementioned questions, we propose simulating patients at the time of admission by extracting admission notes from MIMIC-III discharge summaries. Models built on top of this cohort can act as early warning systems in the absence of key vital signs, thereby assisting clinicians from the very beginning and preventing mistakes. Furthermore, models developed using this cohort can also help hospital management plan resources more effectively by predicting length of stay (LoS). We have also designed ClinicalML, a novel pipeline that extracts key indicators for mortality and length of stay using Named Entity Recognition (NER), followed by a K-Means feature reducer. The extracted features were then trained and tested using a battery of traditional ML models.

2 Related Work

Several methods for mortality prediction and hospital ICU length of stay have been adopted on large datasets, including MIMIC-III and MIMIC-IV [7], as well as private datasets from hospitals and institutions. However, the methods span from using handcrafted features using on traditional ML models to using end-to-end BERT models. Furthermore, each study employs a different variant of the underlying cohort which makes it difficult to assess the quality of the applied approach.

Score based prediction Various score based methods exist to predict mortality. Acute Physiology and Chronic Health Evaluation (APACHE) uses 34 physiological measures captured during the initial 24 hours after ICU admission. Similarly, Simplified Acute Physiology Score (SAPS)[8] uses 12 physiological measures and Quick Sepsis-related Organ Failure Assessment Score (qSOFA)[9] uses 3 physiological measures. APACHE was further improved to include various measures not available at the time of admission such as chronic health variables of AIDS, etc.[10]

Clinical outcome prediction using vitals Vitals contain critical real-time data on heart rate, blood pressure, oxygen saturation, etc. that can be used to train various ML and Deep Learning (DL) models. The AIMS model [11] used vital signs in training a CNN-LSTM model to predict mortality. AIMS model used MIMIC-III data to create different cohorts for 3, 7, and 14- days. AIMS also reports a confidence score which is the difference between boundary probability (= 0.5) and the predictive probability. Chen et. al [12] proposed an attention-based temporal convolution network that uses 48 hours of vitals data to predict mortality. Their model outperformed traditional ML models in terms of F1 score. Sadaghi [10] proposed a battery of ML models for predicting mortality based on the first hour of ICU admission. Alghatani et. al [13] developed a model for length of stay and mortality prediction using 12 baseline demographic features(age, gender, height, weight, etc.), and vitals. The final model can be integrated into the Intelligent Remote Patient Monitoring (IRPM) framework. Alabbad et. al[14] used vitals to predict the length of stay of COVID-19 patients in Saudi Arabia using a battery of ML models.

Clinical outcome prediction using textual data While vitals can provide accurate clinical outcome predictions, in most cases where patient has just been admitted to ICU, critical vitals are unavailable. Hence, various studies have been conducted to utilize unstructured textual data to predict clinical outcomes in the future. Chrusciel et al.[15] extracted known vitals from clinical texts and demographic features to for prediction of length of stay. Aken et. al[3] used BERT based models to predict clinical outcomes using admission notes generated from discharged summaries. Hashir et al[16] used deep learning methods to model the temporal nature of an ICU to predict mortality using unstructured clinical notes. Lee et al[17] proposed a neural network architecture that employs an auxiliary loss to ensure that clinical note embeddings capture diagnostic information effectively.

Clinical outcome prediction using multimodal approach Unstructured data when combined with vitals can further improve the predictive prowess of the underlying models. Yang et al. [18] propose a multimodal model using fusion based approach to integrate time-series vitals data with clinical notes. Yang et al. [19] employ a LSTM model to capture features from time-series vitals data and CNN based model to capture features from clinical notes. Khadanga et al. [20] propose a multi-modal deep neural network that uses recurrent units for time-series data and convolutional networks for clinical notes. Zhu et al. [21] used Retrieval-Augmented Generation (RAG) to extract entities from both time-series data and clinical notes. It aligns these entities with a professional knowledge graph (PrimeKG) for consistency and richer semantics. The framework generates task-relevant summaries of patients' health statuses, which are then fused with other modalities using an adaptive multimodal fusion network.

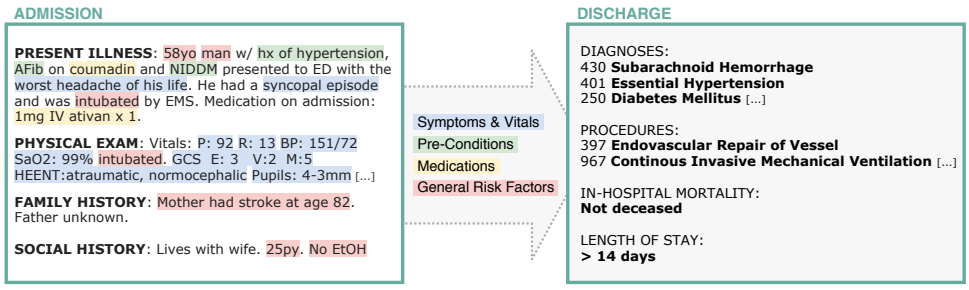

Figure 1: Adapted from [3] *Admission to discharge* sample that demonstrates the outcome prediction task. The model has to extract patient variables and learn complex relations between them in order to predict the clinical outcome.

Single-label tasks: Samples per class					
Mortality		**Length of Stay (in days)**			
0	1	≤ 3	>3 & ≤ 7	>7 & ≤ 14	>14
43,609	5,136	5,596	16,134	13,391	8,488

Table 1: Distribution of labels for *Mortality Prediction* and *Length of Stay* task. Both tasks have unbalanced class distributions.

3 Dataset

Our primary data source is EHR data from MIMIC III v1.4[22] which includes de-identified clinical notes from the Intensive Care Unit (ICU) of Beth Israel Deaconess Medical Center in Massachusetts. For mortality and readmission prediction, we have utilized the cohort established by Van Aken et al. (2021)[3], which simulates newly arrived patients by extracting admission notes from MIMIC III discharge summaries. Figure 1 describes the schematic of generating the cohort.

The cohort filters the discharge summary section provided NOTEEVENTS.csv of MIMIC-III by sections that are known at admission such as: *Chief complaint, (History of) Present illness, Medical history, Admission Medications, Allergies, Physical exam, Family history and Social history.* The task outcome labels like in-hospital mortality and length of stay is extracted from ADMISSIONS.csv table of MIMIC-III.

While preparing the cohort, direct indications of task outcome such as mentions of *patient deceased* for mortality or dates mentioning discharge are filtered to prevent data leak. Moreover, cases of deaths immediately following ICU admission have been omitted. After cleaning, the data amounts to 48,745 and 43,609 training instances for MP and LoS respectively. The data distribution for MP and LoS are provided in Table 1 and patient demographic details in Table 2.

Criteria	**Class**	**Train (MP)**	**Test (MP)**	**Train (LoS)**	**Test (LoS)**
Gender	Male	19102	5482	17190	4941
	Female	14852	4340	13231	3856
Ethnicity	Asian	810	218	718	199
	Black	3137	1047	2896	961
	Hispanic	1184	326	1121	302
	White	24381	7019	21905	6271
	Excluded	4442	1212	3781	1064
Age Group	0-20	310	74	298	72
	20-40	3076	920	2933	876
	40-60	9471	2707	8778	2518
	60-80	14135	4148	12674	3678
	80+	6962	1973	5738	1653

Table 2: MIMIC-III MP and LoS Cohorts

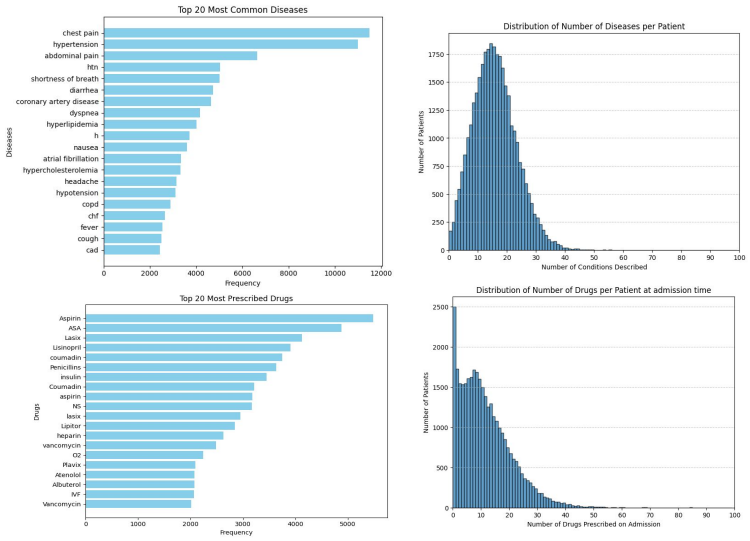

Figure 2: From top-left to bottom-right: a.) 20 Most common diseases found in patients. b.) Distribution of number of diseases among patients c.) 20 Most common therapeutics prescribed. d.) Distribution of number of therapeutics prescribed to patients at admission time

4 Methods

In this study, we hypothesize that both therapeutics and disease entities are critical predictors of clinical outcomes. To test this hypothesis, we develop the *ClinicalML* pipeline, a machine learning framework that consists of three main stages: (1) data preprocessing and feature preparation, (2) feature representation, and (3) classification using traditional machine learning models.

4.1 Feature Extraction and Selection

To identify and extract relevant features from clinical notes, we employ Named Entity Recognition (NER) tools to extract mentions of therapeutics and diseases. For therapeutic entities, we utilize the Med7 NER tagger [23], while for disease entities, we adopt HunFlair [24]. Figure 4b illustrates examples of extracted disease entities, including acronyms such as systolic blood pressure(SBP) and misspelled terms like "in hct."

To address noise and redundancy in the extracted entities, we perform entity normalization by clustering semantically similar terms. Specifically, we apply K-Means clustering on word embeddings generated by BioClinicalBERT [25]. Using the FAISS library [26], we run K-Means with 256 clusters and 200 iterations to obtain cluster centroids. Each extracted entity is then mapped to its corresponding centroid. As shown in Figure 4b, this process groups misspelled terms and abbreviations with their canonical forms into the same cluster.

4.2 Feature Representation

The mapped entities are used to construct a one-hot encoding that captures the presence or absence of each normalized disease entity. This representation is particularly important, as the absence of certain diseases can also significantly impact the model's predictions.

4.3 Training Classifiers

The one-hot encoded feature vectors are used to train a suite of traditional machine learning models, including Logistic Regression (LR), Random Forest (RF), Gradient Boosted Trees (GBBoost), and eXtreme Gradient Boosting (XGBoost) [27, 28, 29, 30].

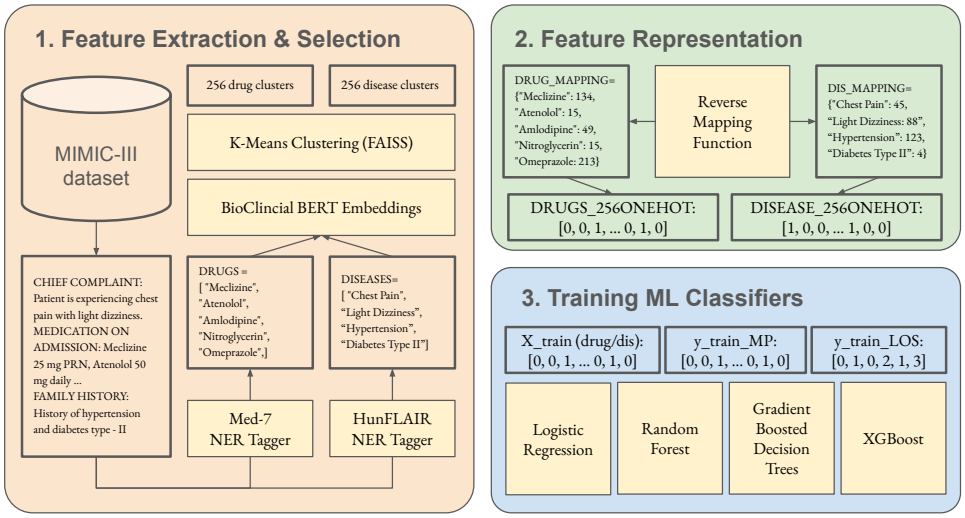

Figure 3: Overview of *ClincialML* pipeline. We first construct the admission notes cohort based on [3]. We then utilize Med7 NER tagger [23] and HunFlair [24] to extract drug and disease names respectively. For dimensionality reduction, we adopt K-means clustering on BioClincicalBERT embeddings of the extracted entites, to derive 256-sized clusters which are then reverse mapped on the entities to create OneHot Vectors. Using reduced OneHot feature vectors of therapeutics and diseases, we train a suite of ML models on the Mortality Prediction and Length of Stay Task.

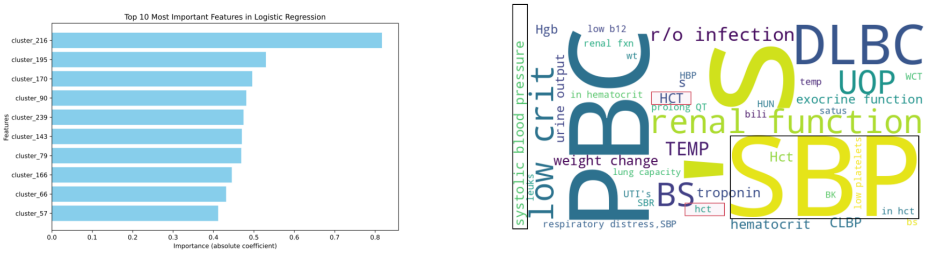

Figure 4: From left to right a.) Feature importance analysis for Logistic Regression for LoS task trained using diseases. Cluster 216 is the most important cluster for LoS task. b.) Word cloud of cluster 216, illustrating all the diseases associated with cluster 216.

Model (Features)	Technique	Precision	Recall	F1-score
XGBoost (Clinical Text)	Imbalanced Classes	0.63	0.51	0.50
	SMOTE	0.56	0.59	0.57
	Oversampling	0.62	0.55	**0.58**
XGBoost (Disease)	Imbalanced Classes	0.60	0.51	0.50
	SMOTE	0.52	0.52	0.52
	Oversampling	0.57	0.64	**0.58**
XGBoost (Therapeutics)	Imbalanced Classes	0.65	0.51	0.50
	SMOTE	0.51	0.52	0.49
	Oversampling	0.56	0.61	**0.56**

Table 3: Comparison of XGBoost performance using different resampling techniques for Mortality. Oversampling results in best performance for all our feature sets (marked in **bold**).

Model (Features)	Technique	Precision	Recall	F1-score
XGBoost (Clinical Text)	Imbalanced Classes	0.35	0.30	0.30
	SMOTE	0.32	0.32	**0.32**
	Oversampling	0.35	0.32	**0.32**
XGBoost (Disease)	Imbalanced Classes	0.34	0.33	0.29
	SMOTE	0.31	0.33	0.32
	Oversampling	0.33	0.36	**0.33**
XGBoost (Therapeutics)	Imbalanced Classes	0.34	0.28	0.25
	SMOTE	0.28	0.29	0.28
	Oversampling	0.30	0.31	**0.30**

Table 4: Comparison of XGBoost performance using different resampling techniques for LoS. Oversampling results in best performance for all our feature sets (marked in **bold**).

5 Experiments

We conducted series of experiments to evaluate the performance of various approaches on our clinical outcome prediction tasks. Each experiment was designed to test different aspects of feature representation and model robustness, especially in handling challenges like class imbalance and high-dimensional data. We baseline our methods against the approach proposed by Van Aken et al.[3] by training and testing BioBERT[31], SciBERT[32], and UMLS-BERT models[33]. For benchmarking, we consider macro F1-score, which averages F1 scores across classes giving equal weightage to all classes irrespective of their proportions in the data. Table 5 and Table 6 depict comprehensive results covering all our experiment combinations.

5.1 Imbalanced Classes: No Resampling vs. SMOTE vs. Oversampling

Class imbalance poses a significant challenge in clinical datasets, particularly for mortality prediction where the majority class can dominate. To address this, we compared the performance of models trained on the original dataset (no resampling) with those using Synthetic Minority Oversampling Technique (SMOTE) [34] and simple oversampling of the minority class to make all instances equal to majority class. These methods aim to mitigate imbalance by either synthesizing new samples (SMOTE) or duplicating existing ones (oversampling). For all types of features, XGBoost model has illustrated that a simple oversampling technique of resampling the underrepresented tasks by duplicating existing examples has outperformed SMOTE and imbalanced dataset. An overall improvement ranges between 0.01 - 0.07 in terms of macro F1 score when using simple oversampling technique. Table 3 and Table 4 demonstrate results obtained using various sampling techniques.

5.2 Using only Therapeutics as Features

For this experiment, we utilized extracted therapeutic-related features. Therapeutic features were extracted using the Med7 entity tagger. This process involved identifying and tagging all mentions of medications in the clinical text, followed by the creation of a vocabulary of unique drugs. Due to the large vocabulary size, we applied dimensionality reduction techniques to ensure computational efficiency. Specifically, we used BioClinicalBERT embeddings to convert the drugs into dense vector representations and performed clustering using FAISS k-Means. This enabled us to group similar drugs into clusters of sizes 256.

The models were then trained using both one hot encoded features and the reduced feature sets. For both MP and LoS task using only therapeutics, XGBoost has achieved an F1-score(macro) of 0.56 and 0.30 respectively.

5.3 Using Only Diseases as Features

Disease mentions in clinical text often serve as stronger predictors for clinical outcomes compared to drug mentions. For this experiment, we extracted disease entities using the Flair Named Entity Recognition (NER) framework. Similar to the approach used for drugs, we applied BioClinicalBERT embeddings to represent diseases and reduced the dimensionality through FAISS k-Means clustering.

Traditional ML models trained using only diseases features were able to outperform the models trained using Therapeutics. For both MP and LoS task XGBoost has outperformed other ML models, XGBoost has achieved an F1-score(macro) of 0.58 and 0.33 respectively. The results also illustrate that disease-based features are more informative for both mortality prediction and length of stay when compared to therapeutics based features.

5.4 Using Doc2Vec Embeddings as Features

To also explore the potential of lightweight models capturing holistic information from the clinical text, we experimented with Doc2Vec [35] embeddings. This approach leverages BioWordVec [36] to generate dense representations of entire clinical documents, allowing the model to learn from the full text without the need for explicit feature engineering. By using this method, we aimed to capture the semantic context and relationships within the text that may not be evident from individual features such as drugs or diseases.

XGBoost trained on Doc2Vec embedding has outperformed other traditional ML models, it has achieved an F1-score(macro) of 0.58 and 0.32 for MP and LoS task respectively. The results demonstrate that Doc2Vec embeddings provide a balanced representation of the clinical text, improving the model's ability to generalize across different tasks.

5.5 Feature Importance Analysis

For the feature importance analysis we have considered Logistic Regression(LR) for length of stay task trained using diseases. Figure 4 illustrates the most important weights LR. The weights associated with cluster 216, which contains diseases such as systolic blood pressure (SBP), troponin, hematocrit (HCT) etc. have shown indications for length of stay.[37, 38, 39, 40] Lower levels of SBP are indicative for longer stay in ICU as it signifies more sever underlying condition.[39] Lower levels of HCT are indicative of longer stay in ICU.[40] Higher levels of Troponin is indicative for longer stay in the ICU as troponin is indicative of heart heart.[37]

6 Discussion and Conclusion

Our proposed method achieves results close to BERT based baselines. XGBoost trained on disease, therapeutics and entire clinical text features yields the best results for both tasks. Mortality prediction in clinical settings is inherently challenging, largely due to the complexity and variability of patient data. Furthermore, our cohort designed at admission time is especially challenging as the amount of information is limited. One key takeaway from our study is the weak correlation between therapeutics mentioned in discharge notes and mortality outcomes. Diseases show stronger predictive power for mortality outcomes as shown by our results. Furthermore, feature importance of LR(Figure 4) for LoS trained on diseases illustrates strong clinical correlations. Our results also demonstrate the value of traditional machine learning (ML) models. Although they slightly underperform compared to BioBERT in terms of macro F1-score, they achieve comparable results within a 10% margin while offering greater explainability. This makes them a more trustworthy option from clinicians point of view scenarios where interpretability is a priority.[41] However, traditional ML models might struggle to capture nuanced relationships between drugs, diseases, and outcomes, which language models like BioBERT can identify more effectively. For instance, our proposed approach is unable to capture the level of SBP i.e. whether it is high or lower. It only captures the presence of SBP. Finally, our analysis highlights the importance of task-specific feature engineering. While feature embeddings like Doc2Vec outperform NER-based pipelines by leveraging holistic information from the clinical text, diseases remain the most reliable predictors. Future work could explore integrating richer temporal data and real-time patient monitoring to further improve clinical prediction models.

Model	Features	Precision	Recall	F1-score
BioBERT	Entire clinical text	0.62	0.72	**0.64**
SciBERT	Entire clinical text	0.63	0.60	0.61
UMLS-BERT	Entire clinical text	0.51	0.50	0.47
XGBoost	Doc2Vec	0.62	0.55	***0.58***
	Diseases	0.57	0.64	***0.58***
	Therapeutics	0.56	0.61	0.56
Gradient Boosted Trees	Doc2Vec	0.55	0.62	0.52
	Diseases	0.57	0.66	***0.56***
	Therapeutics	0.53	0.54	0.53
Logistic Regression	Doc2Vec	0.57	0.69	0.54
	Diseases	0.57	0.67	***0.55***
	Therapeutics	0.55	0.63	0.54
Random Forests	Doc2Vec	0.95	0.50	0.47
	Diseases	0.56	0.51	0.50
	Therapeutics	0.55	0.52	***0.52***

Table 5: Performance metrics of various models and features for mortality prediction. The best F1-score overall is highlighted in **bold**. The best F1-score for each model across the three feature sets is highlighted in *italic* + **bold**.

Model	Features	Precision	Recall	F1-score
BioBERT	Entire clinical text	0.39	0.35	0.36
SciBERT	Entire clinical text	0.43	0.37	**0.38**
UMLS-BERT	Entire clinical text	0.39	0.35	0.35
XGBoost	Doc2Vec	0.35	0.32	0.32
	Diseases	0.33	0.36	***0.33***
	Therapeutics	0.30	0.31	0.30
Gradient Boosted Trees	Doc2Vec	0.30	0.32	0.30
	Diseases	0.33	0.35	***0.31***
	Therapeutics	0.29	0.29	0.29
Logistic Regression	Doc2Vec	0.34	0.38	***0.32***
	Diseases	0.33	0.35	***0.32***
	Therapeutics	0.29	0.31	0.29
Random Forests	Doc2Vec	0.34	0.30	0.30
	Diseases	0.33	0.33	***0.33***
	Therapeutics	0.30	0.30	0.30

Table 6: Performance comparison of models with different features for length of stay (LoS) prediction. The best F1-score overall is highlighted in **bold**. The best F1-score for each model across the three feature sets is highlighted in *italic* + **bold**.

References

[1] Rajeev Bopche, Lise Tuset Gustad, Jan Egil Afset, Birgitta Ehrnström, Jan Kristian Damås, and Øystein Nytrø. In-hospital mortality, readmission, and prolonged length of stay risk prediction leveraging historical electronic patient records. *JAMIA Open*, 7(3):ooae074, 09 2024.

[2] Roy Adams, Katharine E. Henry, Anirudh Sridharan, Hossein Soleimani, Andong Zhan, Nishi Rawat, Lauren Johnson, David N. Hager, Sara E. Cosgrove, Andrew Markowski, Eili Y. Klein, Edward S. Chen, Mustapha O. Saheed, Maureen Henley, Sheila Miranda, Katrina Houston, Robert C. Linton, Anushree R. Ahluwalia, Albert W. Wu, and Suchi Saria. Prospective, multi-site study of patient outcomes after implementation of the trews machine learning-based early warning system for sepsis. *Nature Medicine*, 28(7):1455–1460, July 2022.

[3] Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online, April 2021. Association for Computational Linguistics.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[6] Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for bert, 2021.

[7] Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1–10, 2023.

[8] J R Le Gall, P Loirat, A Alperovitch, P Glaser, C Granthil, D Mathieu, P Mercier, R Thomas, and D Villers. A simplified acute physiology score for ICU patients. *Crit. Care Med.*, 12:975–977, 1984.

[9] Matthew M Churpek, Ashley Snyder, Xuan Han, Sarah Sokol, Natasha Pettit, Michael D Howell, and Dana P Edelson. Quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores for detecting clinical deterioration in infected patients outside the intensive care unit. *Am. J. Respir. Crit. Care Med.*, 195:906–911, 2017.

[10] William Romine Reza Sadeghi, Tanvi Banerjee. Early hospital mortality prediction using vital signals, 2019.

[11] Ian Atkinson Stephanie Baker, Wei Xiang. Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach - Scientific Reports nature.com. `https://www.nature.com/articles/s41598-020-78184-7`, 12 2020.

[12] Yu-wen Chen, Yu-jie Li, Peng Deng, Zhi-yong Yang, Kun-hua Zhong, Li-ge Zhang, Yang Chen, Hong-yu Zhi, Xiao-yan Hu, Jian-teng Gu, Jiao-lin Ning, Kai-zhi Lu, Ju Zhang, Zheng-yuan Xia, Xiao-lin Qin, and Bin Yi. Learning to predict in-hospital mortality risk in the intensive care unit with attention-based temporal convolution network. *BMC Anesthesiology*, 22(1):119, Apr 2022.

[13] Khalid Alghatani, Nariman Ammar, Abdelmounaam Rezgui, and Arash Shaban-Nejad. Predicting intensive care unit length of stay and mortality using patient vital signs: Machine learning model development and validation. *JMIR Med. Inform.*, 9(5):e21347, May 2021.

[14] Dina A. Alabbad, Abdullah M. Almuhaideb, Shikah J. Alsunaidi, Kawther S. Alqudaihi, Fatimah A. Alamoudi, Maha K. Alhobaishi, Naimah A. Alaqeel, and Mohammed S. Alshahrani. Machine learning model for predicting the length of stay in the intensive care unit for covid-19 patients in the eastern province of saudi arabia. *Informatics in Medicine Unlocked*, 30:100937, 2022.

[15] Jan Chrusciel, François Girardon, Lucien Roquette, David Laplanche, Antoine Duclos, and Stéphane Sanchez. The prediction of hospital length of stay using unstructured data. *BMC Medical Informatics and Decision Making*, 21(1):351, Dec 2021.

[16] Mohammad Hashir and Rapinder Sawhney. Towards unstructured mortality prediction with free-text clinical notes. *Journal of Biomedical Informatics*, 108:103489, 2020.

[17] Sanghoon Lee, Gwanghoon Jang, Chanhwi Kim, Sejeong Park, Kiwoong Yoo, Jihye Kim, Sunkyu Kim, and Jaewoo Kang. Enhancing clinical outcome predictions through auxiliary loss and sentence-level self-attention. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1210–1217, 2023.

[18] Bo Yang and Lijun Wu. How to leverage multimodal ehr data for better medical predictions?, 2021.

[19] Haiyang Yang, Li Kuang, and FengQiang Xia. Multimodal temporal-clinical note network for mortality prediction. *Journal of Biomedical Semantics*, 12(1):3, Feb 2021.

[20] Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. Using clinical notes with time series data for ICU management. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6432–6437, Hong Kong, China, November 2019. Association for Computational Linguistics.

[21] Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. Emerge: Integrating rag for improved multimodal ehr predictive modeling, 2024.

[22] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May 2016.

[23] Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. Med7: a transferable clinical natural language processing model for electronic health records, 2020.

[24] Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. Hunflair: An easy-to-use tool for state-of-the-art biomedical named entity recognition, 2020.

[25] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings, 2019.

[26] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[27] R. E. Wright. Logistic regression, 1995.

[28] Leo Breiman. Random forests, 2001.

[29] Jerome H. Friedman. Stochastic gradient boosting, 2002.

[30] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, August 2016.

[31] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.

[32] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text, 2019.

[33] George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus, 2021.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002.

[35] Quoc V. Le and Tomás Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014.

[36] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific Data*, 6(1):52, May 2019.

[37] Adam J Singer, Samita Heslin, Hal Skopicki, Chen On, Lisa B Senzel, Mathew Tharakan, Henry C Thode, Jr, and Frank Peacock. Introduction of a high sensitivity troponin reduces ED length of stay. *Am. J. Emerg. Med.*, 76:82–86, February 2024.

[38] Jian Guan, Michael Karsy, Meic H Schmidt, and Erica F Bisson. Impact of preoperative hematocrit level on length of stay after surgery on the lumbar spine. *Global Spine J.*, 5(5):391–395, October 2015.

[39] J D Chalmers, A Singanayagam, and A T Hill. Systolic blood pressure is superior to other haemodynamic predictors of outcome in community acquired pneumonia. *Thorax*, 63(8):698–702, August 2008.

[40] Mehmet Toptas, Nilay Sengul Samanci, İbrahim Akkoc, Esma Yucetas, Egemen Cebeci, Oznur Sen, Mehmet Mustafa Can, and Savas Ozturk. Factors affecting the length of stay in the intensive care unit: Our clinical experience. *Biomed Res. Int.*, 2018:9438046, March 2018.

[41] Qian Xu, Wenzhao Xie, Bolin Liao, Chao Hu, Lu Qin, Zhengzijin Yang, Huan Xiong, Yi Lyu, Yue Zhou, and Aijing Luo. Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review. *J. Healthc. Eng.*, 2023(1):9919269, February 2023.