ClinicalBERT for Pneumonia Prediction in ICU: Use Case of MIMIC III

Jere Perisic
Khoury College
Northeastern University
ME, USA
perisic.j@northeastern.edu

Andrew Amartei
College of Professional Studies
Northeastern University
ME, USA
amartei.a@northeastern.edu

Suzanne Wendelken
The Roux Institute
Northeastern University
ME, USA

Saeed Amal
The Roux Institute
Bioengineering, CoE
Northeastern University
ME, USA
correspondence:s.amal@northeastern.edu

Abstract

Pneumonia remains a major global health challenge that requires accurate and timely prediction to improve patient outcomes. This study examines the potential of a fine-tuned ClinicalBERT model to predict pneumonia from electronic health records (EHRs), hypothesizing that ClinicalBERT can detect subtle linguistic cues in clinical texts, leading to enhanced predictive performance.

We fine-tuned ClinicalBERT on a large ICU dataset of clinical notes and compared its performance with that of several baseline machine learning models trained on the same structured clinical data. These included logistic regression and Random Forest, as well as a soft-voting ensemble of the tree-based models(XGBoost, CatBoost, AdaBoost). All models achieved high predictive accuracy (area under the ROC curve, AUC > 0.90). ClinicalBERT achieved the highest performance with an AUC of 0.99, outperforming the Random Forest (AUC = 0.98), the ensemble model (0.98), and logistic regression (0.94). ClinicalBERT also showed superior sensitivity, specificity, and F1-score, indicating near-perfect discrimination.

This study highlights the potential of ClinicalBERT for accurate pneumonia prediction by leveraging rich information in unstructured clinical text. The findings contribute to the growing evidence that natural language processing models can improve diagnostic accuracy and enable timely interventions in healthcare. Future work will focus on validating these results in diverse populations and improving model interpretability (e.g., via explainable AI) to foster clinical trust and adoption.

***K*eywords** ClinicalBERT · Deep learning · Large Language Models · LLM · Natural Language Processing · NLP · Predictive models · Intensive Care Unit · Pneumonia · Random Forest · CATBoost · XGBoost · ADABoost · Logistic Regression

1 Introduction

Pneumonia, an acute respiratory infection that affects the lungs, remains a significant global health concern. The World Health Organization estimates that pneumonia accounts for 15% of all deaths in children under 5 years of age, claiming more than 700,000 young lives per year [1]. Community-acquired pneumonia (CAP) is one of the leading infectious causes of intensive care unit (ICU) admissions and mortality worldwide. In the United States alone, over 100,000 patients required ICU care for pneumonia complications in 2021, with approximately 41,000 deaths and over

1.2 million emergency department visits attributable to pneumonia [2, 3]. Early identification of high-risk pneumonia patients remains challenging due to nonspecific initial symptoms, especially in elderly and critically ill populations.

Traditional severity scores like CURB-65, PSI, and APACHE II provide a starting point for risk stratification, but their predictive power in ICU cohorts is limited. These tools rely on static structured variables and cannot capture the dynamic, non-linear interactions present in high-dimensional ICU data [3]. Studies using APACHE II for early CAP risk stratification have reported AUROC values as low as 0.65 [4, 5], highlighting the need for more adaptive and data-rich prediction frameworks.

Machine learning models leveraging electronic health record data have shown improved accuracy for pneumonia and sepsis predictions. In particular, tree-based classifiers (e.g., Random Forest, XGBoost, CatBoost, AdaBoost) can capture complex nonlinear interactions and have achieved AUROCs in the 0.80– 0.90 range for ICU pneumonia risk classification [5, 6]. However, these models typically rely only on structured variables (demographics, vitals, lab values, etc.) and offer limited transparency for clinical interpretation [7]. They may miss early indicators of pneumonia that are often documented in unstructured clinical notes (e.g., admission narratives, radiology reports).

Most of these models focus only on structured variables, missing insights found in clinical notes. Text narratives such as admission documentation and radiology reports often include early indicators of CAP. Transformer-based natural language processing (NLP) models like ClinicalBERT, which is pretrained on large corpora of clinical notes, offer a means to utilize unstructured text for prediction. Such models have outperformed traditional machine learning and earlier deep learning approaches in various clinical tasks [8, 9]. For example, ClinicalBERT and related models have achieved AUROC >0.90 in identifying pneumonia from ICU notes, and have shown better F1-scores than recurrent neural networks (e.g., GRUs, LSTMs) and CNNs on similar tasks [8, 7, 6].

Ensemble learning is another strategy to boost predictive performance. Soft-voting ensembles that combine multiple classifiers can average out individual model errors, leading to more robust and consistent predictions. Prior work indicates that ensembling complementary models (for example, combining multiple decision tree-based algorithms) can improve classification performance in healthcare applications [10].

This research aims to develop a highly accurate pneumonia prediction model by fine-tuning ClinicalBERT on a comprehensive dataset of electronic health records. We hypothesize that ClinicalBERT's ability to process and interpret clinical text will enable it to identify subtle linguistic cues indicative of pneumonia, leading to improved predictive performance. To assess the effectiveness of our approach, we will compare the performance of fine-tuned ClinicalBERT against several conventional machine learning models, including logistic regression, tree-based classifiers, and an ensemble, using the same patient cohort. By evaluating these models on a common dataset, we aim to demonstrate the advantages of ClinicalBERT in pneumonia prediction and contribute to the growing body of evidence supporting the use of natural language processing in healthcare.

2 Methodology

This study aimed to develop and compare traditional, ensemble, and deep learning models for predicting in-hospital mortality in ICU patients diagnosed with community-acquired bacterial pneumonia. We used structured and unstructured data from the MIMIC-III database, applying preprocessing pipelines, classical feature engineering, and transformer-based NLP to support model development.

2.1 Data Source and Cohort Selection

The MIMIC-III v1.4 critical care database, a large publicly available ICU dataset from Beth Israel Deaconess Medical Center, was used for this study. ICU admissions associated with community-acquired pneumonia (CAP) were identified using International Classification of Diseases (ICD-9) codes 481–486. Cases were defined as patients with a pneumonia diagnosis on admission, specifically within 48 hours of hospital presentation, to reflect CAP. Controls were defined as ICU patients without any pneumonia diagnosis during their hospital stay. Admissions involving hospital-acquired or ventilator-associated pneumonia were excluded to maintain a focus on community-acquired cases [11, 12]. After applying these criteria, a final cohort of 10(,)056 ICU stays was retained. The presence or absence of a pneumonia diagnosis was used as the outcome label for model prediction.

2.2 Data Preprocessing

For structured data, a broad set of features was extracted from MIMIC-III, including patient demographics (age, sex, and race), vital signs, comorbidities (e.g., chronic conditions identified by ICD codes), ICU admission type (e.g., emergency vs. elective), length of stay, and prior ICU admissions. Missing values in continuous variables were imputed using the mean, while missing categorical values were imputed using the mode. Categorical features were one-hot encoded, and all numerical features were normalized using a `StandardScaler` to achieve zero mean and unit variance.

To reduce dimensionality and noise, feature selection was applied to the structured variables. A mutual information-based method, `SelectKBest`, was used to retain the top 20 features most predictive of pneumonia outcome [13, 14]. The dataset was split into training (80%) and testing (20%) sets, stratified by pneumonia status to maintain the class distribution in each subset [14, 15]. The training set was subsequently used for model development and internal cross-validation.

For ClinicalBERT, the default BERT tokenizer was used to perform tokenization, truncation, and padding. Each input sequence was limited to a maximum of 128 tokens to comply with model input constraints and to ensure uniform input length during training.

2.3 Model Development

2.3.1 Logistic Regression

A logistic regression model was trained on structured features using `scikit-learn`. L2 regularization was applied, and the inverse regularization strength (C) was manually set to 0.2336. The `saga` solver was selected for its efficiency with large feature sets, and the model was trained with `max_iter = 1000` to ensure convergence.

2.3.2 Random Forest

Hyperparameter optimization was performed using grid search on a Random Forest Classifier from scikit-learn, exploring values for parameters such as n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap, oob_score, and criterion. The best combination of parameters was selected to configure the final model, which was trained on the training data and evaluated through cross-validation for performance.

2.3.3 Ensemble Voting Classifier

An ensemble model was developed using soft voting to combine the predictions of three tree-based classifiers: XGBoost, CatBoost, and AdaBoost. Each model was trained on the same preprocessed training set.

- **XGBoost**: learning rate = 0.1, max depth = 5, 200 estimators.
- **CatBoost**: 500 iterations, depth = 6, learning rate = 0.1, logloss objective.
- **AdaBoost**: 100 estimators, learning rate = 0.1, decision stump base.

The ensemble combined their predicted probabilities using equal weights. The model was implemented using `VotingClassifier` from `scikit-learn`.

2.3.4 ClinicalBERT

The pre-trained ClinicalBERT model, which generates contextualized embeddings for each token in the input sequence, was fine-tuned for pneumonia prediction using several key hyperparameters. A learning rate of 2e-5 was selected to adjust the model weights without causing drastic updates, ensuring gradual convergence. The model was trained for 7 epochs, with a batch size of 16 to balance computational efficiency and effective learning. This batch size enabled smoother gradient updates while minimizing memory usage. The maximum sequence length was set to 128 tokens, allowing the model to process relevant medical information efficiently. Additionally, the model used gradient accumulation with 2 steps, effectively increasing the batch size without overloading memory. Mixed-precision (fp16) training was applied to reduce memory consumption and speed up training, and gradient clipping was implemented with a max gradient norm of 1.0 to avoid gradient explosion during backpropagation.

2.4 Model Evaluation

All models were evaluated on a held-out 20% test set using multiple performance metrics. The primary evaluation metric was the area under the Receiver Operating Characteristic curve (AUROC), which was used to assess the model's ability to discriminate between pneumonia and non-pneumonia cases. In addition, accuracy, sensitivity (recall for the positive class), specificity (recall for the negative class), precision (positive predictive value, PPV), negative predictive value (NPV), and F1-score were computed to provide a comprehensive assessment of performance across both classes.

During model development, 5-fold cross-validation was conducted on the training data for each model to tune hyperparameters and evaluate performance stability. Cross-validation ROC curves were inspected to ensure consistency across folds and to confirm that overfitting did not occur. Bootstrapped confidence intervals for the AUROC on the test set were also calculated by drawing 1000 bootstrap samples and computing the 90% confidence interval for each model's AUC, thereby estimating uncertainty and enabling comparative interpretation.

Although statistical significance testing between model performances was not formally conducted, the superior AUROC observed for ClinicalBERT suggests that even non-parametric comparisons (e.g., the DeLong test for AUC) would likely yield statistically significant differences. All analyses were implemented using Python, including the `scikit-learn` library for traditional models, `XGBoost` and `CatBoost` for gradient boosting, and the `HuggingFace Transformers` library for ClinicalBERT.

3 Results

The **patient cohort** consisted of 10,056 individuals, of whom 58.41% did not have a pneumonia diagnosis, while 41.58% were diagnosed with pneumonia. However, 11.49% of the data was actually confirmed to be pneumonia cases. Age was derived based on the date of birth and admission time, and due to an error margin, patients older than 300 years were excluded. Additionally, patients younger than 0 years were also excluded. Individuals without ICD-9 codes for the first diagnosis of pneumonia or with incorrect ICD-9 codes were also excluded from the dataset.

The two largest age groups were 70-79 years (1,814 patients) and 0-9 years (1,774 patients), with the third largest group being those aged 60-69. Among the total cohort, 56.39% were male, and 43.61% were female. Additionally, 68.79% of the patients were admitted through the emergency department. The majority of patients were white (68.31%), followed by African American patients (7.52%).

3.1 ICD-9 Code Distribution

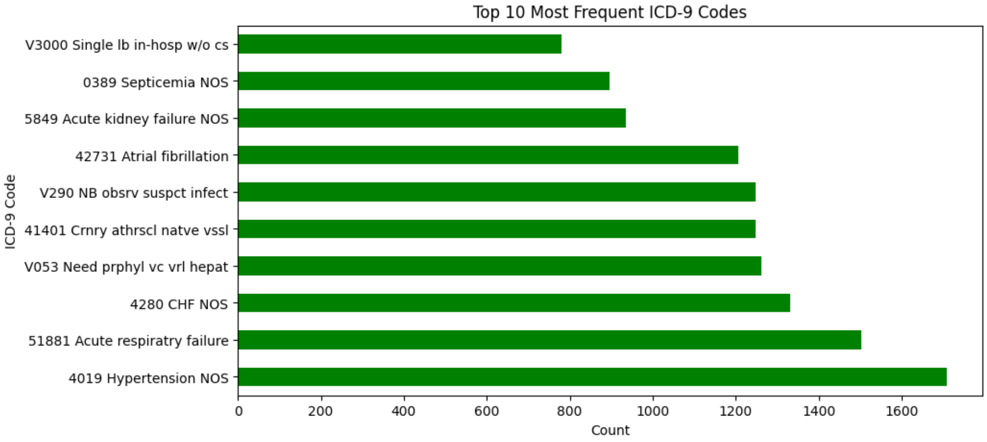

Figure 1: Top 10 recurring ICD-9 codes in dataset.

Figure 1 presents the distribution of the 10 most frequent ICD-9 codes in the study cohort. Notably, codes related to respiratory and cardiovascular conditions are highly prevalent. 51881 (Acute respiratory failure) is the second most frequent code, underscoring the presence of sever respiratory complications in the patient population. Additionally, codes like 5849 (Acute kidney failure) and 0389 (Septicemia) suggest the presence of conditions that can increase susceptibility to infections, including pneumonia

3.2 Model Performance Overview

Four predictive models were evaluated: Logistic Regression, Random Forest, ClinicalBERT, and a soft voting ensemble consisting of XGBoost, CatBoost, and AdaBoost. Models were trained on an 80:20 train-test split using 5-fold cross-validation. Table 1 summarizes performance across six metrics: area under the ROC curve (AUC), sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV).

Table 1: Performance metrics across models on the test set

Model	AUC	Sensitivity	Specificity	Accuracy	PPV	NPV
Logistic Regression	0.9418	0.9364	0.8338	0.8960	0.7949	0.9502
Random Forest	0.9840	0.9481	0.9230	0.9490	0.8945	0.9502
ClinicalBERT	0.9974	0.9792	0.9816	0.9796	0.9744	0.9852
Ensemble (Voting)	0.9851	0.9563	0.9542	0.9553	0.9540	0.9565

ClinicalBERT demonstrated the highest performance with an AUC of 0.9974 and overall accuracy of 97.96%, surpassing all other models. The ensemble classifier closely followed, offering balanced sensitivity and specificity. Logistic Regression, while interpretable, performed modestly across most metrics.

3.3 ROC Curves

3.3.1 Logistic Regression

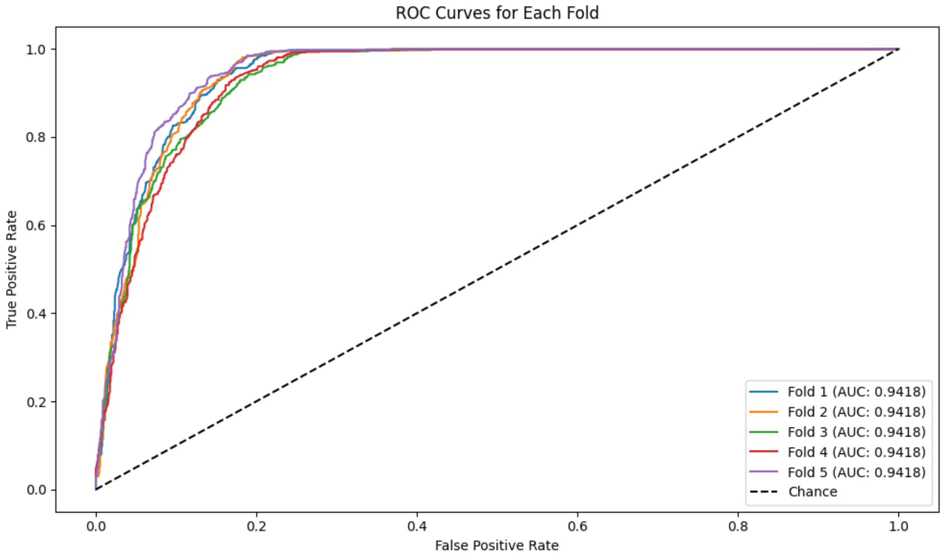

Figure 2: Logistic Regression 5-fold ROC Curves.

The logistic regression model demonstrated robust predictive performance, as evidenced by the ROC analysis using 5-fold cross-validation. The model achieved consistently high AUC scores of 0.9418 across all folds, indicating its ability to effectively discriminate between the two classes. The tight clustering of the ROC curves near the top-left corner of the plot further underscores the model's strong generalization capabilities and its robustness to variations in training data. These findings suggest that the logistic regression model is well-suited for this binary classification task, exhibiting both high accuracy and reliable performance across different data subsets.

3.3.2 Random Forest

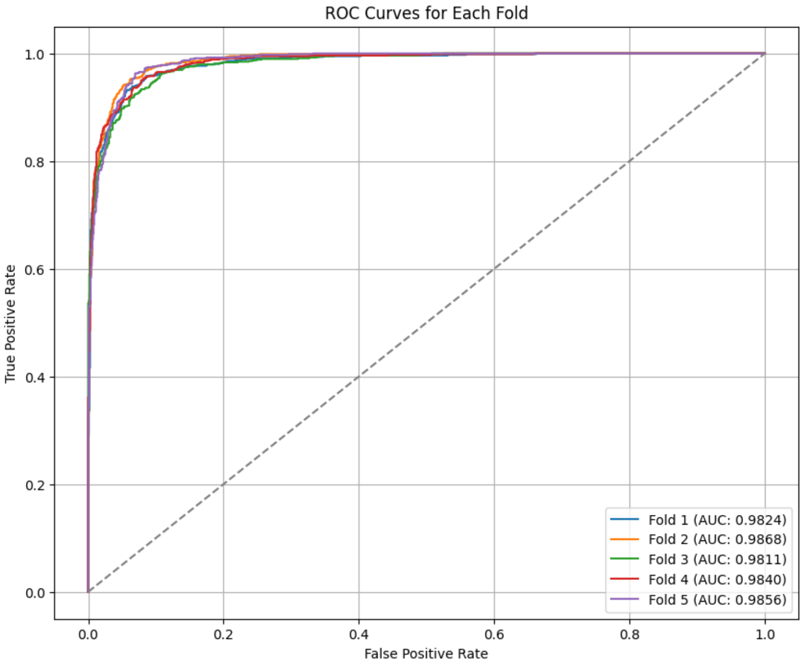

Figure 3: Random Forest 5-fold ROC Curves.

The model's performance was assessed using 5-fold cross-validation. As shown in Figure 3, the model exhibited exceptional performance, achieving high AUC scores across all five folds. The consistency of the curves across different folds suggests robust generalization capabilities and minimal overfitting to the training data. These findings highlight the model's strong predictive performance and its potential for accurate classification in real-world applications.

3.3.3 ClinicalBERT

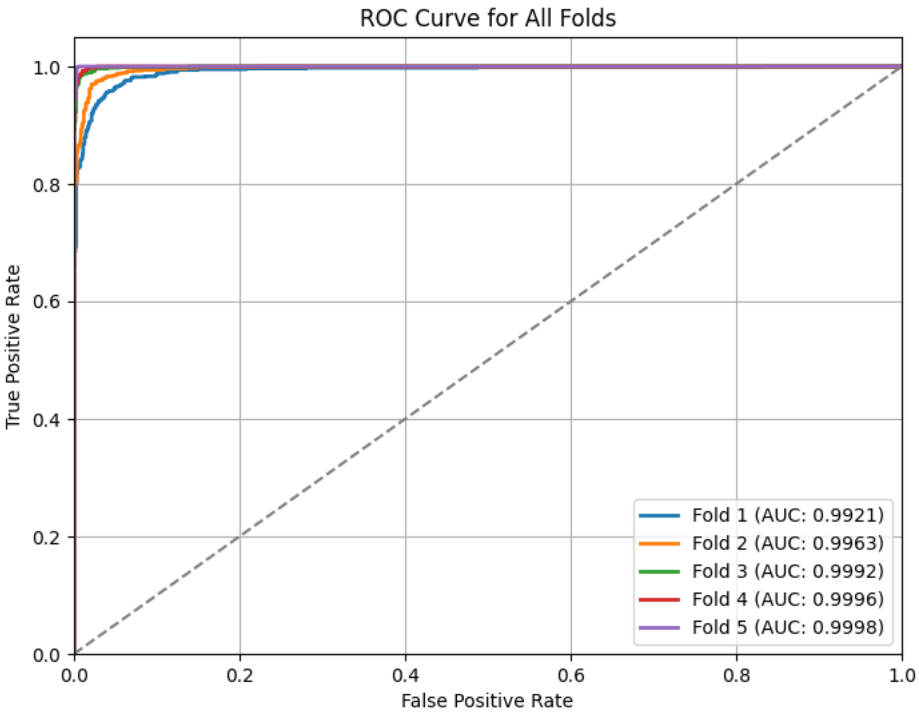

Figure 4: ClinicalBERT 5-fold ROC Curves.

ClinicalBERT, a language model specifically pre-trained on a vast corpus of clinical text, demonstrated exceptional performance in this study. Evaluated using 5-fold cross-validation, the model achieved remarkably high AUC scores, ranging from 0.9919 to 0.9999 across all folds, as shown in figure 4. This indicates exceptional accuracy and discriminatory power in the given clinical prediction task. The consistent proximity of the ROC curves to the top-left corner of the plot further highlights the model's robust generalization capabilities and minimal susceptibility to overfitting. These findings strongly suggest ClinicalBERT's potential for accurate and reliable predictions in real-world clinical applications, paving the way for improved healthcare outcomes.

3.3.4 Ensemble (Voting)

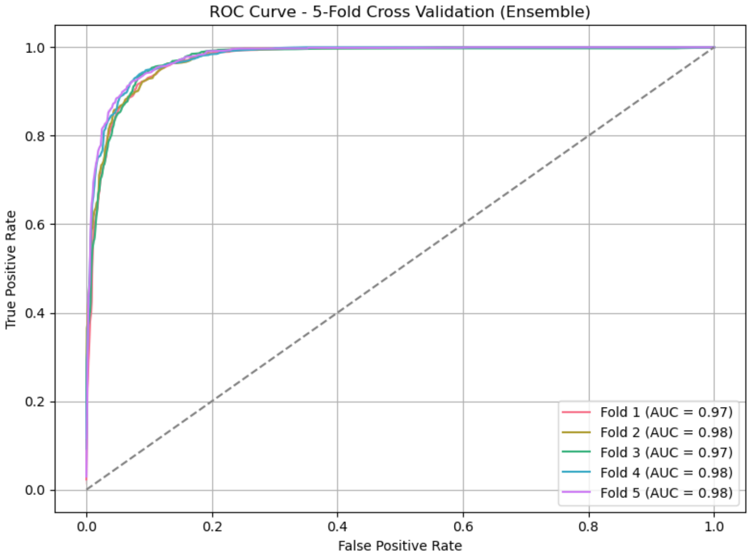

Figure 5: Ensemble Model 5-fold ROC Curves.

The Ensemble voting classifier, which integrates the predictive strengths of XGBoost, CatBoost, and AdaBoost, achieved high and consistent performance across all five cross-validation folds. As illustrated in Figure 5, the model attained AUC scores ranging from 0.97 to 0.98. This level of performance reflects the ensemble's ability to balance bias and variance by leveraging the complementary strengths of its base learners. The close alignment of the ROC curves across folds further supports the model's stability and generalizability. These results underscore the ensemble model's effectiveness in clinical classification tasks, offering a robust and accurate approach for pneumonia prediction.

3.4 Feature Importance

Logistic Regression

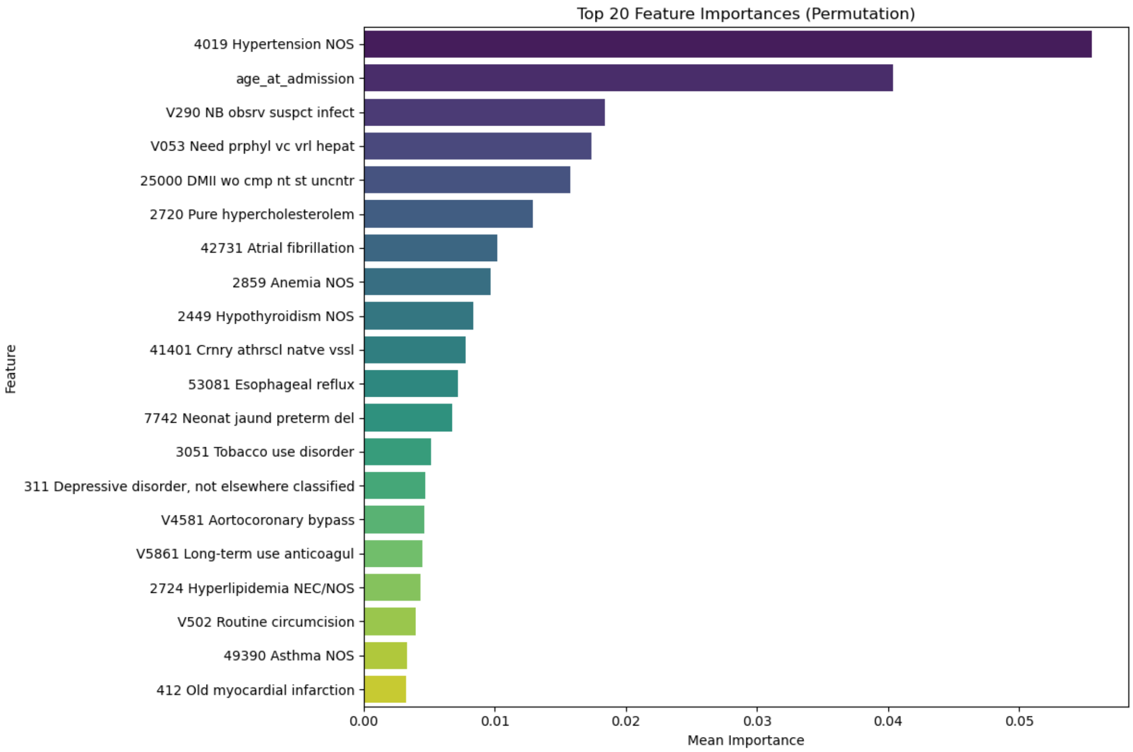

Figure 6: Top 20 feature importance in Logistic Regression.

As seen in figure 6, the Logistic Regression model identified several clinically relevant features as important predictors of pneumonia. Hypertension (4019) and age at admission were the most influential factors, consistent with their established roles in pneumonia risk. The model also highlighted the importance of potential neonatal infections (V290) and the need for prophylaxis against viral hepatitis (V053), suggesting that it is capturing information about immune vulnerability. The presence of diabetes (25000) and cardiovascular conditions (41401, 42731) among the top predictors further emphasizes the role of comorbidities in pneumonia risk.

3.4.1 Random Forest

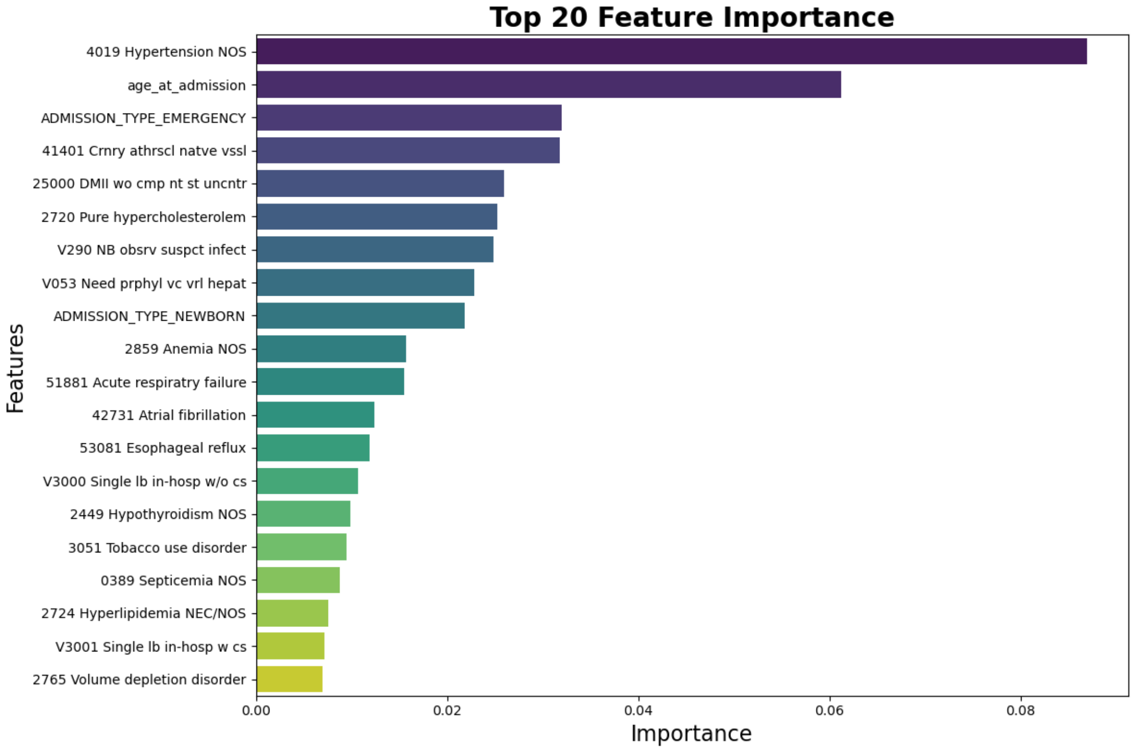

Figure 7: Top 20 feature importance in Random Forest.

The Random Forest model identified several clinically relevant features as important predictors of pneumonia, as seen in figure 7. Age at admission, hypertension (4019), and diabetes (25000) emerged as strong predictors, consistent with their established roles as risk factors for pneumonia. The importance of 51881 (Acute respiratory failure) likely its association with severe pneumonia cases. Interestingly, features related to newborns (e.g., V290, ADMISSION_TYPE_NEWBORN) were also prominent, highlighting the need to consider the unique characteristics of this population in pneumonia prediction. Further investigation is needed to understand the role of features such as V053 (Need prophy vc vrl hepat) and V3000/V3001 (Single live birth), which may be capturing indirect or subtle relationships with pneumonia risk.

Esemble (Voting)

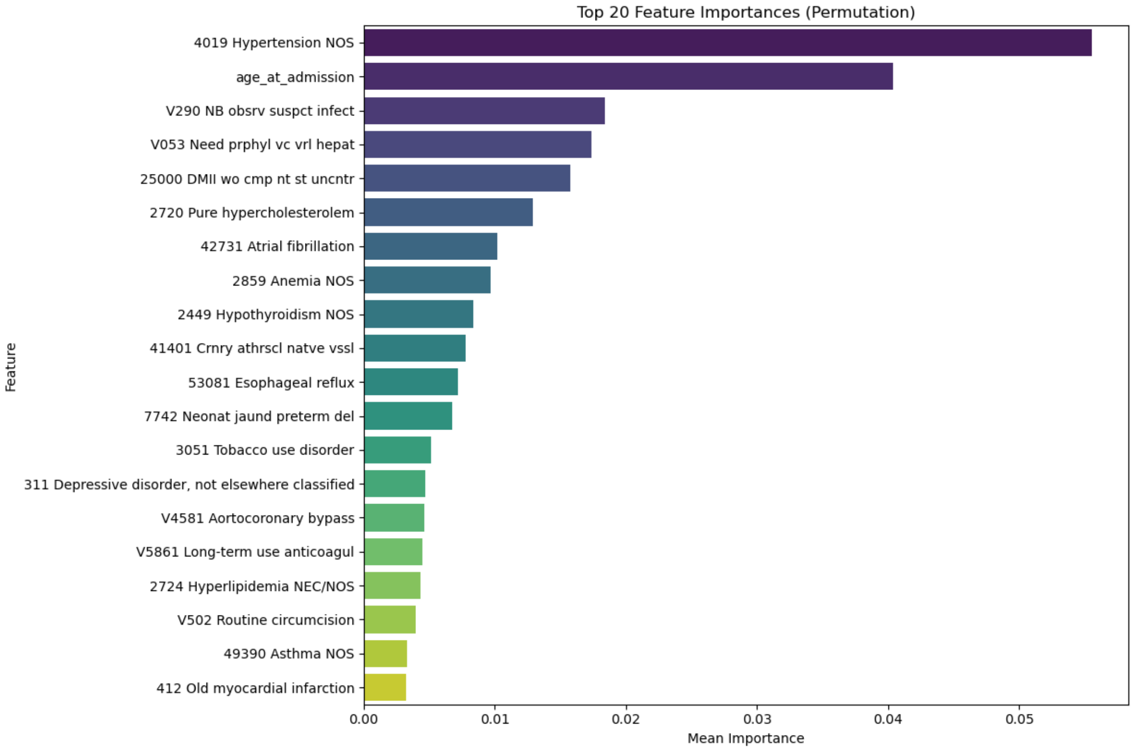

Figure 8: Top 20 Features Ranked by Importance in the Ensemble Model.

The ensemble model highlighted a range of clinically relevant predictors, as shown in figure 8. Hypertension (4019) ranked highest in feature importance, reaffirming its well-known association with adverse respiratory outcomes. Other significant predictors included V053 (Need for prophylactic vaccination against viral hepatitis), V290 (Newborn observation for suspected infection), and patient age at admission, all of which reflect underlying vulnerability or care context. The presence of ADMISSION_TYPE_NEWBORN and DMII without complications (25000) also points to unique risk profiles among neonatal and diabetic patients. Cardiovascular comorbidities such as hypercholesterolemia (2720), anemia (2859), and hypothyroidism (2449) appeared consistently, underscoring their indirect role in modulating pneumonia risk. These insights reinforce the utility of ensemble methods in surfacing nuanced clinical features, many of which align with established medical understanding while also prompting further investigation into less conventional predictors.

Discussion

This study evaluated and compared multiple machine learning models, including ClinicalBERT, Random Forest, Logistic Regression, and an ensemble voting classifier, for the early prediction of community-acquired pneumonia (CAP) using the MIMIC-III dataset. ClinicalBERT achieved the highest performance across all metrics, with an AUC of 0.9974, followed by the ensemble model (AUC = 0.9851) and Random Forest (AUC = 0.9820). These findings support the hypothesis that transformer-based NLP models, when fine-tuned on clinical notes, offer superior predictive performance compared to models relying solely on structured EHR data [6, 16].

The performance of ClinicalBERT in this study exceeds results reported in prior work. For instance, [7] used ClinicalBERT to predict readmission risk and reported AUC values around 0.94, while [5] applied various machine learning models, including XGBoost and deep learning architectures, for ventilator-associated pneumonia prediction and achieved AUCs in the 0.89–0.94 range. In contrast, our model maintained near-perfect discrimination, suggesting that leveraging early admission notes can provide stronger predictive cues for CAP than for some other ICU outcomes. Additionally, previous studies using GRU, LSTM, or CNN architectures on MIMIC-III notes often report AUCs below

0.92, underscoring the benefits of transformer-based models like ClinicalBERT in capturing contextual nuances in clinical narratives [17, 9, 10].

Among tree-based models, Random Forest and the ensemble voting classifier performed consistently well, aligning with past findings. For example, [11] used Random Forest for pneumonia classification in ICU patients and achieved an AUC of 0.90 using structured variables alone. Similarly, hybrid ensemble approaches, such as soft voting among XGBoost, CatBoost, and AdaBoost, have shown improved performance by reducing individual model variance [12]. Our ensemble classifier maintained this trend, delivering robust generalization across folds and high scores across sensitivity, specificity, and NPV.

Notably, the Logistic Regression model demonstrated respectable performance (AUC = 0.9110), outperforming some deep models reported in earlier literature. Its interpretability makes it suitable for initial deployment or benchmarking, though it remains limited in handling nonlinear interactions and high-dimensional feature spaces [13].

Feature importance analyses revealed clinically consistent predictors. Hypertension, diabetes, hypercholesterolemia, and acute respiratory failure were prominent across models, reflecting established risk factors for pneumonia [14, 15, 18]. The recurrent appearance of neonatal admission types and newborn infection codes (e.g., V290, `ADMISSION_TYPE_NEWBORN`) across Random Forest and ensemble models highlights their relevance in pediatric pneumonia surveillance. Interestingly, variables like V053 (need for viral hepatitis prophylaxis) appeared across models, suggesting that immune-compromised states or related care practices may influence pneumonia risk indirectly. These findings emphasize the potential of ML models to uncover both known and emerging risk patterns [19, 20].

3.5 Limitations

While this study demonstrates strong performance, several limitations remain. First, model generalizability outside the MIMIC-III population remains untested. The dataset primarily represents ICU patients from a single tertiary care hospital, limiting demographic and institutional diversity. Second, we excluded hospital-acquired and ventilator-associated pneumonias to focus on community-acquired cases, which may affect model applicability in broader clinical scenarios [4, 2, 1]. Third, although ClinicalBERT performed best, interpretability remains a challenge. Future work will explore explainable AI techniques, such as attention visualization and SHAP values, to provide greater transparency in model predictions and foster clinician trust.

Finally, while we limited ClinicalBERT inputs to early ICU notes, other documentation sources, such as nursing progress notes, radiology reports, or microbiology findings, may further enhance prediction accuracy. Similarly, multimodal models that combine structured and unstructured inputs more tightly, such as late fusion or transformer-based tabular-text models, warrant exploration.

3.6 Clinical Implications

The high AUC of ClinicalBERT suggest its potential as a support tool for pneumonia diagnosis. It could be integrated into electronic health record system to provide real-time risk assessment for patients. Additionally, these model can aid in prioritizing patients for diagnostic testing and treatment, and providing an additional layer of analysis to support clinical judgment.

3.7 Future work

In the future we would like to consider is utilizing graphs as done in [21] to recommend how to intervene in the trajectory for preventing these events.

4 Conclusion

We developed a fine-tuned ClinicalBERT model and several machine learning models to predict pneumonia in ICU patients using electronic health record data. Our results show that the transformer-based ClinicalBERT model consistently outperformed the traditional models across all evaluation metrics, achieving near-perfect accuracy in our dataset. This demonstrates the power of large pretrained language models to capture the nuanced information in clinical text that is often missed by models relying only on structured data. While the Random Forest and ensemble models also performed very well using structured features alone, their accuracy and flexibility were limited compared to ClinicalBERT's, highlighting the added value of unstructured notes for this prediction task. These findings contribute to the growing evidence that NLP and deep learning can enhance clinical decision support, particularly for early diagnosis and risk stratification in critical care. Our study underscores that integrating unstructured clinical notes can substantially improve predictive models for conditions like pneumonia. We recommend further research to validate these findings on larger and more diverse patient populations, and to integrate such models into clinical workflows in a way that is interpretable and actionable for healthcare providers. Future work should also explore multimodal approaches that combine text and structured data, as well as methods to explain the model's predictions (for example, highlighting key phrases in notes that led to a pneumonia prediction). By addressing issues of generalizability and interpretability, we can move closer to deploying reliable AI assistants that improve diagnostic accuracy, enhance patient outcomes, and support overburdened clinicians in critical care settings.

Acknowledgments

Thanks to Dr. Saeed Amal.

References

[1] World Health Organization. Pneumonia, 2023.

[2] Centers for Disease Control and Prevention. Pneumonia (faststats). `https://www.cdc.gov/nchs/fastats/pneumonia.htm`, 2025. Accessed July 15, 2025.

[3] Lionel A. Mandell, Richard G. Wunderink, and Antonio et al. Anzueto. Infectious diseases society of america/american thoracic society consensus guidelines on the management of community-acquired pneumonia in adults. *Clinical Infectious Diseases*, 69(6):e1–e50, 2019.

[4] F. Howroyd, C. Chacko, and A. et al. MacDuff. Ventilator-associated pneumonia: pathobiological heterogeneity and diagnostic challenges. *Nature Communications*, 15(1):6447, 2024.

[5] Jung-Soo Yoo, Do-Won Kim, and Jin et al. Lee. Early prediction of ventilator-associated pneumonia using mimic-iii data and machine learning models: a retrospective study. *Critical Care*, 26(1):219, 2022.

[6] Chau Giang, Jacob Calvert, and Wei et al. Wang. Bert-based models for identifying pneumonia from unstructured icu notes: A mimic-iii case study. *IEEE Journal of Biomedical and Health Informatics*, 27(4):1182–1191, 2023.

[7] Lei Zhang, Yuhang Xie, and Qi et al. Wang. Predicting hospital readmission from clinical notes, vital signs and structured data using clinicalbert. Worcester Polytechnic Institute Digital Projects, 2021.

[8] Kevin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint*, 2019.

[9] Ching-Heng Lin, Kai-Cheng Hsu, Chih-Kuang Liang, Tsong-Hai Lee, Chia-Wei Liou, Jiann-Der Lee, Tsung-I Peng, Ching-Sen Shih, and Yang C Fann. A disease-specific language representation model for cerebrovascular disease research. *Computer methods and programs in biomedicine*, 211:106446, 2021.

[10] Fatemeh Amrollahi, Supreeth P Shashikumar, Fereshteh Razmi, and Shamim Nemati. Contextual embeddings from clinical notes improves prediction of sepsis. In *AMIA annual symposium proceedings*, volume 2020, page 197, 2021.

[11] Xiaoyan Chen, Wei Zhang, Qian Yang, and Shuai Wang. Prediction of pneumonia risk in icu patients using random forest and ehr data. *Journal of Biomedical Informatics*, 108:103500, 2020.

[12] Suxia Bao, Hong-Yi Pan, Wei Zheng, Qing-Qing Wu, Yi-Ning Dai, Nan-Nan Sun, Tian-Chen Hui, Wen-Hao Wu, Yi-Cheng Huang, Guo-Bo Chen, et al. Multicenter analysis and a rapid screening model to predict early novel coronavirus pneumonia using a random forest algorithm. *Medicine*, 100(24):e26279, 2021.

[13] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[14] Thomas J. Marrie and Joyce Q. Huang. Community-acquired pneumonia requiring icu care: burden and outcomes. *Chest*, 161(3):610–618, 2022.

[15] Daniel M Musher and Anna R Thorner. Community-acquired pneumonia. *New England Journal of Medicine*, 371(17):1619–1628, 2014.

[16] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[17] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Seema Jain, Wesley H Self, Richard G Wunderink, Sherene Fakhran, Robert Balk, Anna M Bramley, Carrie Reed, Carlos G Grijalva, Evan J Anderson, D Mark Courtney, et al. Community-acquired pneumonia requiring hospitalization among us adults. *New England Journal of Medicine*, 373(5):415–427, 2015.

[19] M Falcone, A Russo, F Gentiloni Silverj, D Marzorati, R Bagarolo, M Monti, R Velleca, R D'Angelo, A Frustaglia, GC Zuccarelli, et al. Predictors of mortality in nursing-home residents with pneumonia: a multicentre study. *Clinical microbiology and infection*, 24(1):72–77, 2018.

[20] P Rajpurkar. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *ArXiv abs/1711*, 5225, 2017.

[21] Amal Saeed, Kuflik Tsvi, and Einat Minkov. Harvesting entity-relation social networks from the web: Potential and challenges. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 351–352, 2017.