CLINICALBERT IN PNEUMONIA PREDICTION: ACHIEVING HIGH ACCURACY WITH FINE-TUNING

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

ABSTRACT

Pneumonia remains a significant global health challenge, demanding accurate and timely prediction to improve patient outcomes. This study investigates the potential of fine-tuned ClinicalBERT, a state-of-the-art language model, for pneumonia prediction using 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 enhanced predictive performance.

We fine-tuned ClinicalBERT on a comprehensive dataset of clinical notes and compared its performance with Random Forest and Logistic Regression models trained on the same data. Our results demonstrate that all three models achieved high predictive accuracy, with AUC values exceeding 0.9. Notably, ClinicalBERT attained an AUC of 0.99, surpassing both Random Forest (AUC = 0.98) and Logistic Regression (AUC = 0.94)

This study highlights ClinicalBERT's potential for accurate pneumonia prediction by leveraging the rich information embedded in clinical text. The findings contribute to the growing body of evidence supporting the use of natural language processing in healthcare, particularly for enhancing diagnostic accuracy and facilitating timely interventions. Future research will explore the generalization of these findings to diverse patient populations and investigate the interpretability of ClinicalBERT's predictions to foster clinical trust and adoption.

***K*eywords** ClinicalBERT · Pneumonia · Random Forest · 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]. Even in developed countries with advanced healthcare systems, pneumonia poses a serious threat, leading to hospitalization and mortality, particularly among the elderly and individuals with underlying health conditions. [2] The diagnosis of pneumonia can be challenging due to the variability in its presentation, with symptoms often overlapping other respiratory diseases. Early and accurate detection of pneumonia is crucial for prompt initiation of treatment and improved patient outcomes. [3]

Traditional methods for pneumonia diagnosis are based heavily on clinical assessment, chest radiographs, and laboratory tests. However, these approaches can be time-consuming, require specialized expertise, and may not always be definitive in the early stages of the disease. In recent years, machine learning has emerged as a promising tool for improving healthcare care, particularly in disease prediction. Machine learning algorithms can analyze vast amounts of data, identify complex patterns, and rapidly generate predictive models. The ability of these algorithms to leverage diverse data sources, including clinical notes, laboratory results, and imaging data, offers the potential to revolutionize the prediction of pneumonia.[4]

ClinicalBERT, a powerful language model pre-trained on a massive corpus of clinical text, has demonstrated exceptional capabilities in understanding medical terminology and extracting meaningful insights from clinical narratives.[5] By capturing the nuances of human language in medical records, ClinicalBERT can uncover hidden relationships between clinical features and disease outcomes. Previous studies have successfully applied ClinicalBERT to various healthcare tasks, including predicting hospital readmission [6] and classifying diseases.

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 with two established machine learning models: Random Forest and Logistic Regression. 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 Literature Review

2.1 Pneumonia Prediction in the Era of ClinicalBERT

Pneumonia, an acute respiratory infection affecting the lungs, remains a leading cause of morbidity and mortality worldwide. Accurate and timely diagnosis is crucial for effective treatment and improved patient outcomes. This literature review examines existing pneumonia prediction approaches, focusing on the merging role of natural language processing (NLP) and language models like ClinicalBERT.

2.2 Traditional Approaches to Pneumonia Prediction

Traditional methods for pneumonia diagnosis rely heavily on clinical assessment, chest radiographs, and laboratory tests. However, these approaches can be time-consuming, require specialized expertise, and may not always be definitive in the early stages of disease. In recent years, machine learning has emerged as a promising tool for enhancing pneumonia prediction by leveraging diverse data sources, including patient demographics, vital signs, laboratory results, and imaging data. Studies have explored various machine learning models for pneumonia prediction, including:

- **Logistic Regression:** A classical statistical method widely used to identify risk actors associated with pneumonia and predict the likelihood of developing the disease. For instance, logistic regression was employed for predicting pneumonia mortality in nursing home residents [7].
- **Random Forest:** This model has shown promising results in pneumonia prediction due to its ability to handle high-dimensional data and capture complex non-linear relationships. Random forest was developed for early prediction of COVID-19 pneumonia (AUC = 0.956). [8]
- **Deep Learning:** Deep learning models, particularly convolutional neural networks (CNNs), have been successfully applied to pneumonia prediction using chest X-ray images. While outside the scope of this review focused on NLP, their strong performance in identifying pneumonia patterns and predicting disease severity is worth noting. [9]

While these approaches have demonstrated the potential of machine learning for pneumonia prediction, they often rely on structured data and may not fully capture the rich clinical information embedded within unstructured EHR data.

2.3 NLP in Healthcare

NLP has become a powerful tool for extracting meaningful insights from unstructured clinical text. The development of advanced language models, particularly those based on transformer architecture like BERT [10], has revolutionized the field of clinical NLP.

ClinicalBERT [5], a language model pre-trained on a large corpus of clinical notes, has demonstrated state-of-the-art capabilities in understanding medical terminology and extracting relevant information from clinical narratives.

- **Hospital Readmission Prediction:** ClincialBERT demonstrated effectiveness in predicting hospital readmission, highlighting its ability to capture patient-specific information from clinical notes. [6]
- **Disease Classification:** ClinicalBERT has been used to classify various diseases, including heart failure [11], and sepsis[12], demonstrating its potential for accurate and timely diagnosis.

2.4 ClinicalBERT for Pneumonia Prediction

Despite extensive research on ClinicalBERT in healthcare, its specific application to pneumonia prediction remains relatively unexplored. This presents a significant opportunity to leverage the model's ability to process and interpret the clinical text to identify subtle linguistic cues indicative of pneumonia, potentially leading to enhanced predictive performance.

3 Methodology

3.1 Data Source

This study utilized the publicly available MIMIC-III database. Data points without ICD-9 codes (after excluding those corresponding to the targeted diagnosis) and those with incorrectly assigned ICD-9 codes were removed. Pneumonia cases were identified using specific ICD-9 codes as defined in the database, with pneumonia cases matching the regular expression PNA_ICD_REGEX: Final = re.compile(r"^48[0–6]∗$"). Patients without these codes were classified as controls.

In the data preprocessing phase, ICD-9 codes were processed using regular expressions to identify patients with pneumonia, resulting in a binary feature indicating the presence or absence of pneumonia. For each patient, a history of ICD-9 codes was constructed up to the point of their first pneumonia diagnosis. This history was stored as a list of codes for each individual patient. Additionally, patient demographics were extracted and merged with the cohort data. Key demographic details such as patient admission type, ethnicity, and age were also extracted and included in the dataset.

3.2 ICD-9 Code Transformation

- For Random Forest and Logistic Regression code histories were converted into a suitable format using one-hot encoding. This created a binary feature for each unique ICD-9 code in the dataset.
- For ClinicalBERT, the raw ICD-9 codes in the history were tokenized using AutoTokenizer.

3.3 Model Training

The dataset was randomly divided into five folds for 5-fold cross-validation. In each iteration, four folds were used for training, and the remaining fold was used for testing. This process was repeated 5 times, with each fold being tested once.

3.3.1 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.

3.3.2 Logistic Regression

A Logistic Regression model was trained using scikit-learn in Python with key hyperparameters to ensure optimal performance. The max_iter was set to 1000 to allow the model to converge, while random_state was fixed at 13 for reproducibility. The solver was set to 'saga', suited for large datasets, and L2 regularization was applied with the penalty parameter. The C parameter, controlling regularization strength, was set to 0.23357214690901212 to balance underfitting and overfitting. These settings ensured effective training and model generalization.

3.3.3 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.

3.4 Evaluation Metrics

- **Area Under the Receiver Operating Characteristic Curve (AUC):** To measure the model's ability to discriminate between pneumonia and non-pneumonia cases.
- **Sensitivity:** The proportion of true pneumonia cases correctly identified.

$$Sensitivity = \frac{TP}{TP + FN}$$

- **Specificity:** The proportion of true non-pneumonia cases correctly identified.

$$Specificity = \frac{TN}{TN + FP}$$

- **Positive Predictive Value (PPV):** The proportion of predicted pneumonia cases that were actually pneumonia.

$$PPV = \frac{TP}{TP + FP}$$

- **Negative Predictive Value (NPV):** The proportion of predicted non-pneumonia cases that were actually non-pneumonia.

$$NPV = \frac{TN}{TN + FN}$$

- **Confidence Intervals:** 95% confidence intervals were calculated for each metric to estimate the variability of the model's performance.

4 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%).

4.1 ICD-9 Code Distribution

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

4.2 Model Performance

Model	AUC	Sensitivity	Specificity	PPV	NPV
Random Forest	0.9840	0.9481	0.9230	0.8945	0.9628
Logistic Regression	0.9418	0.9364	0.8338	0.7949	0.9502
ClinicalBERT	**0.9974**	**0.9792**	**0.9816**	**0.9744**	**0.9852**

Table 1: Performance of different pneumonia prediction models.

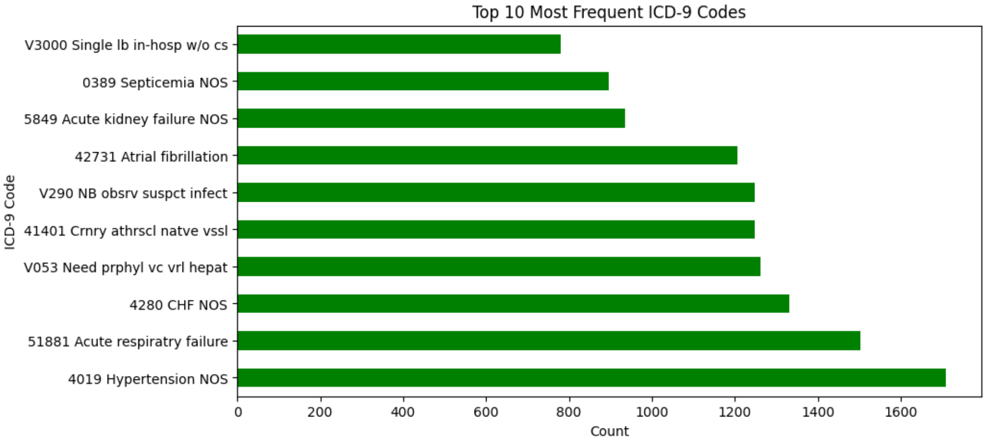

Figure 1: Top 10 recurring ICD-9 codes in dataset.

The table 1 presents a comparative analysis of the performance of the three pneumonia prediction models. While all models demonstrate good classification ability, ClinicalBERT outperforms both Random Forest and Logistic Regression across all evaluated metrics. Its near-perfect AUC (0.9974) and high sensitivity (0.9792) and specificity (0.9816) suggest exceptional accuracy in identifying pneumonia cases. These findings highlight the potential of ClinicalBERT to enhance pneumonia diagnosis and improve patient outcomes. However, further research is needed to address its limitations regarding generalization.

4.3 ROC Curves

4.3.1 Random Forest

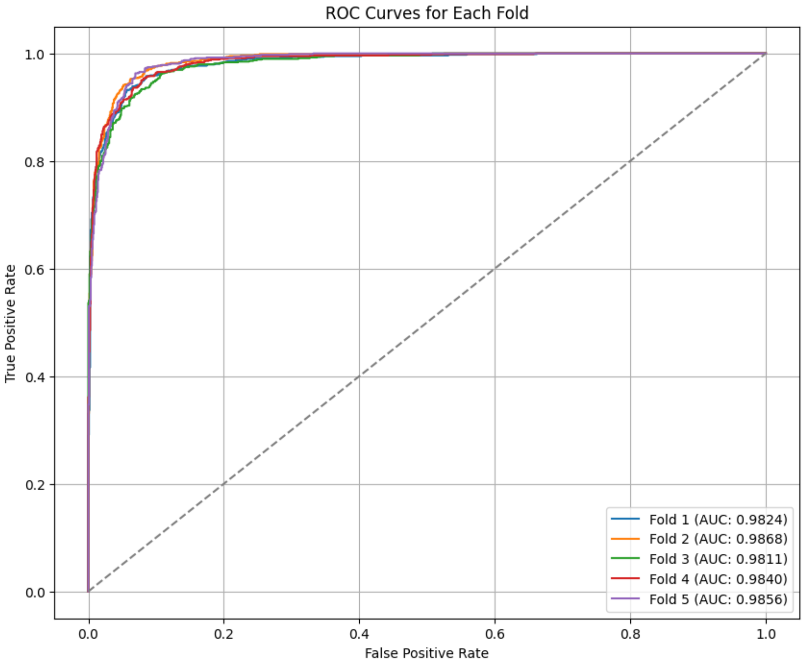

Figure 2: Random Forest 5-fold ROC Curves.

The model's performance was assessed using 5-fold cross-validation. As shown in figure 2, the model exhibited exceptional performance, achieving high AUC scores across all five folds. The consistency of the curves across different folds suggest 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.

4.3.2 Logistic Regression

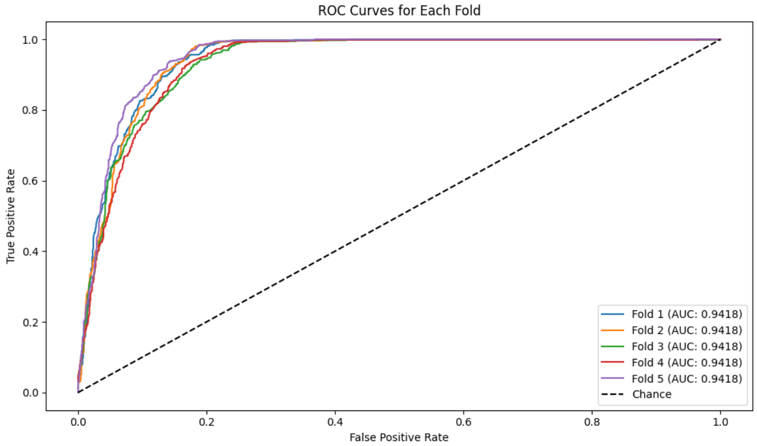

Figure 3: Logistic Regression 5-fold ROC Curves.

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.

4.3.3 ClinicalBERT

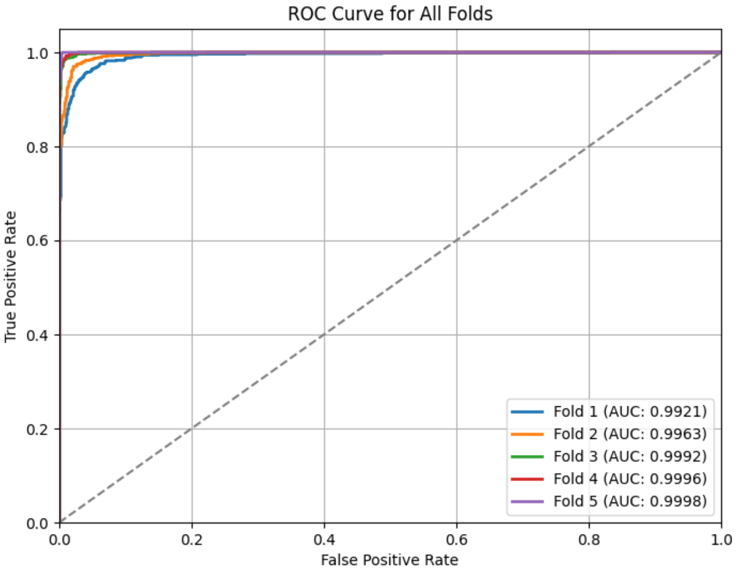

Figure 4: ClinicalBERT 5-fold ROC Curves.

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.

4.4 Feature Importance

4.4.1 Random Forest

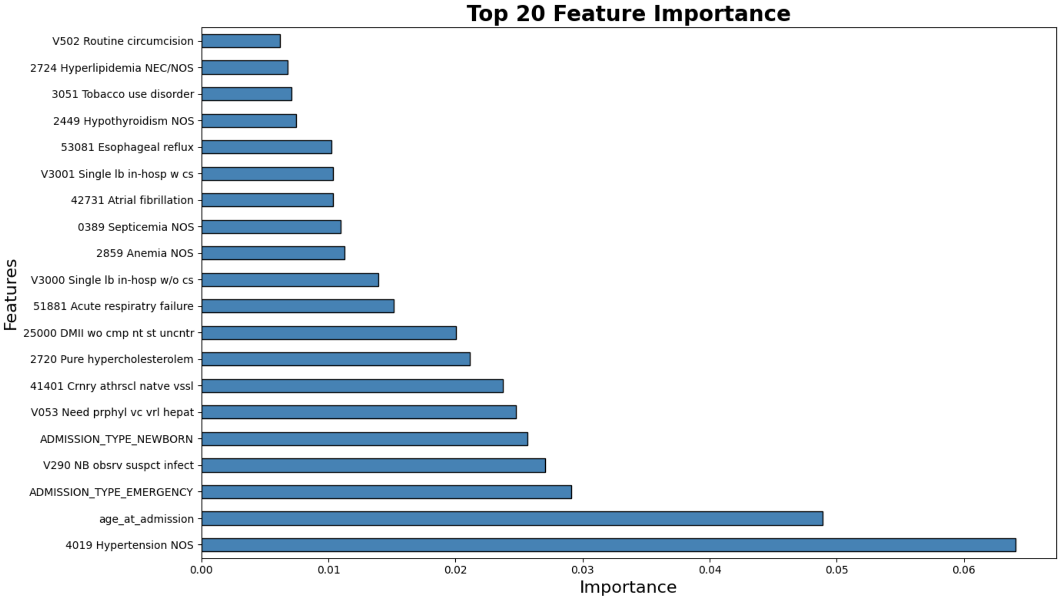

Figure 5: Top 20 feature importance in Random Forest.

The Random Forest model identified several clinically relevant features as important predictors of pneumonia, as seen in figure 5. 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.

4.4.2 Logistic Regression

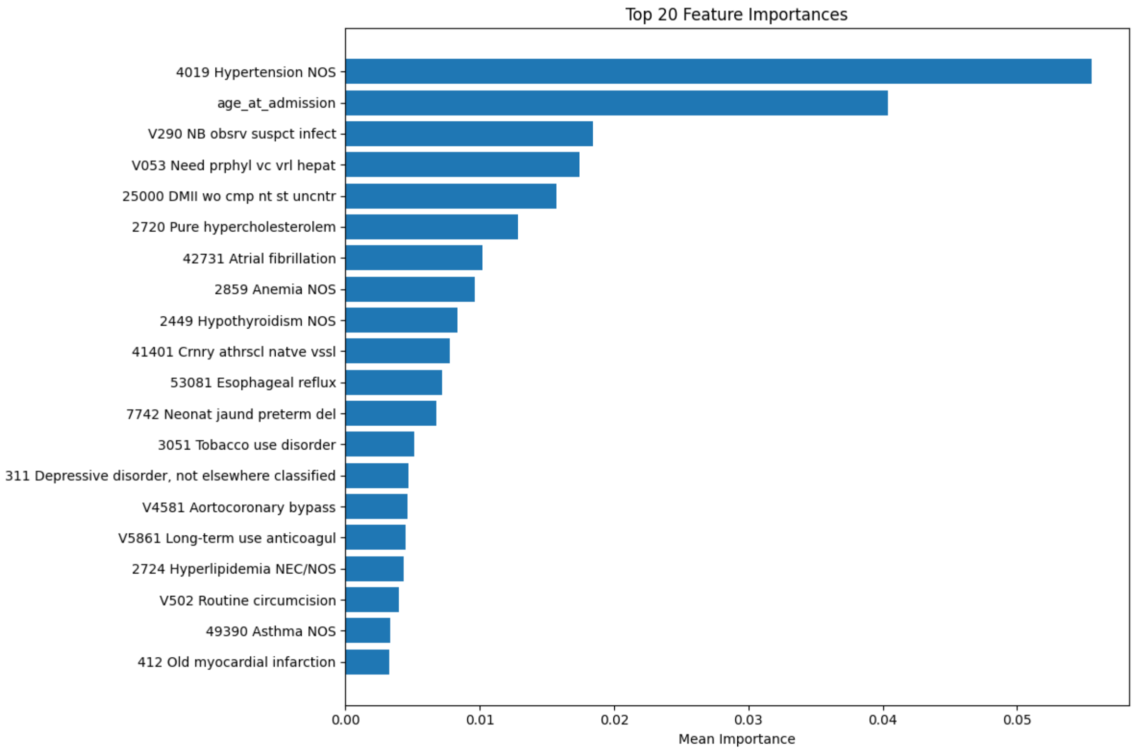

Figure 6: Top 20 feature importance in Logistic Regression.

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.

5 Discussion

This study aimed to evaluate the performance of three machine learning models in predicting pneumonia from electronic health records using ICD-9 codes. Out findings demonstrate superior performance of ClinicalBERT, compared to Random Forest and Logistic Regression.

The superior performance of ClinicalBERT can be attributed to its ability to capture complex relationships and contextual information which traditonal models often miss. While Random Forest and Logistic Regression demonstrated good predictive ability, their reliance on discrete features may limit their ability to capture the full complexity of clinical data.

Each model in this study has its own set of strengths and weaknesses. ClinicalBERT excels in accuracy and capturing complex relationships in clinical text, but it can be difficult to interpret, prone to overfitting, and requires significant computational resources. Random Forest offers good performance, handles high dimensionality well, and provides insights into feature importance, but it can also overfit and may not capture complex interactions as effectively as deep learning models. Logistic Regression, on the other hand, is simple, interpretable, and efficient, but its ability to capture complex non-linear relationships in data is limited.

5.1 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.

5.2 Limitations

This study is limited by its design and reliance on data from single institution, this may limit the generalization. Furthermore, we need to employ explainable AI techniques to understand ClinicalBERT's decision-making process, explore alternative methods for pneumonia identification, such as natural language processing of clinical notes, and testing the models on diverse datasets from multiple institutions.

6 Conclusion

This study investigated the performance of three machine learning models, ClinicalBERT, Random Forest, and Logistic Regression, in predicting pneumonia from electronic health records. ClinicalBERT consistently outpreformed the other two models across all evaluation metrics, demonstrating the potential of transformer-based language models for accurately capturing the complexities of clinical data and predicting pneumonia. While Random Forest and Logistic Regression showed good predictive capability, their performance was limited compared to ClinicalBERT.

Further research should focus on validating these findings on larger, and more diverse datasets from multiple institutions. Exploring alternative methods for pneumonia identification, such as natural language processing of clinical notes, could address potential biases introduced by relying solely on ICD-9 codes. Furthermore, enhancing the interpretability of ClinicalBERT's predictions through explainable AI techniques is crucial for clinical acceptance and trust.

Acknowledgments

Thanks to Dr. Saeed Amal.

References

[1] World Health Organization. Pneumonia, 2023.

[2] 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.

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

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

[5] Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642, 2023.

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

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.