AI-Driven Pneumonia Prediction: A Data-Centric Approach Using Machine Learning

Daniel Akama Nyamweya
SAT 5114

Problem Statement

- Pneumonia remains a leading cause of morbidity and mortality worldwide, making early detection and prediction is crucial for improving patient outcomes.
- This research explores the application of artificial intelligence (AI) in pneumonia prediction using structured clinical datasets.
- By leveraging machine learning techniques, the study aims to develop a predictive model capable of identifying pneumonia risk based on patient demographics, symptoms, laboratory results, and other relevant features.

Literature Review

- *Effah et al. (2022)*: Used 8 ML models on 535 patients with 45 features (e.g., CRP, procalcitonin). Random Forest and XGBoost achieved 92.0% and 90.8% accuracy.

- *Luo et al. (2020)*: Focused on post-kidney transplant patients (n=519), identifying 43 severe pneumonia cases. Random Forest yielded AUROC 0.91 and PPV 0.85. Key features: pulmonary lesions, reoperation.

- *Jeon et al. (2023)*: Predicted in-ICU mortality in pneumonia patients using ML models. Identified risk factors and supported better clinical decisions.

Literature review summary

- *Swetha et al. (2021)*: Used CNNs and ResNet-50 on 26,684 chest X-rays to detect pneumonia early.

- *Bhattarai et al. (2023)*: Developed a CNN-based system for automatic pneumonia detection in chest X-rays. Improved diagnostic speed and accuracy.

Our Approach: Dimensionality Reduction + Predictive Modeling

- **Goal**: Build a predictive model for pneumonia using patient data.
- **Challenge**: The dataset has 70+ features, which can lead to noise and overfitting.
- **Solution**: Applied PCA to reduce dimensionality while preserving variance.
- **Model**: Training base models i.e. Logistic regression and Ensemble models.

Code Highlights: Before and after applying PCA to dataset

```
[5]: # lets print summary infomation on the dataset
print('Summary infomation on dataset')
print('-----------------------------------------------')
describe_df.data_info()
```

```
Summary infomation on dataset
-----------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 83 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   record_id           768 non-null    int64
 1   age                 768 non-null    object
 2   gender              767 non-null    float64
 3   height              407 non-null    object
```

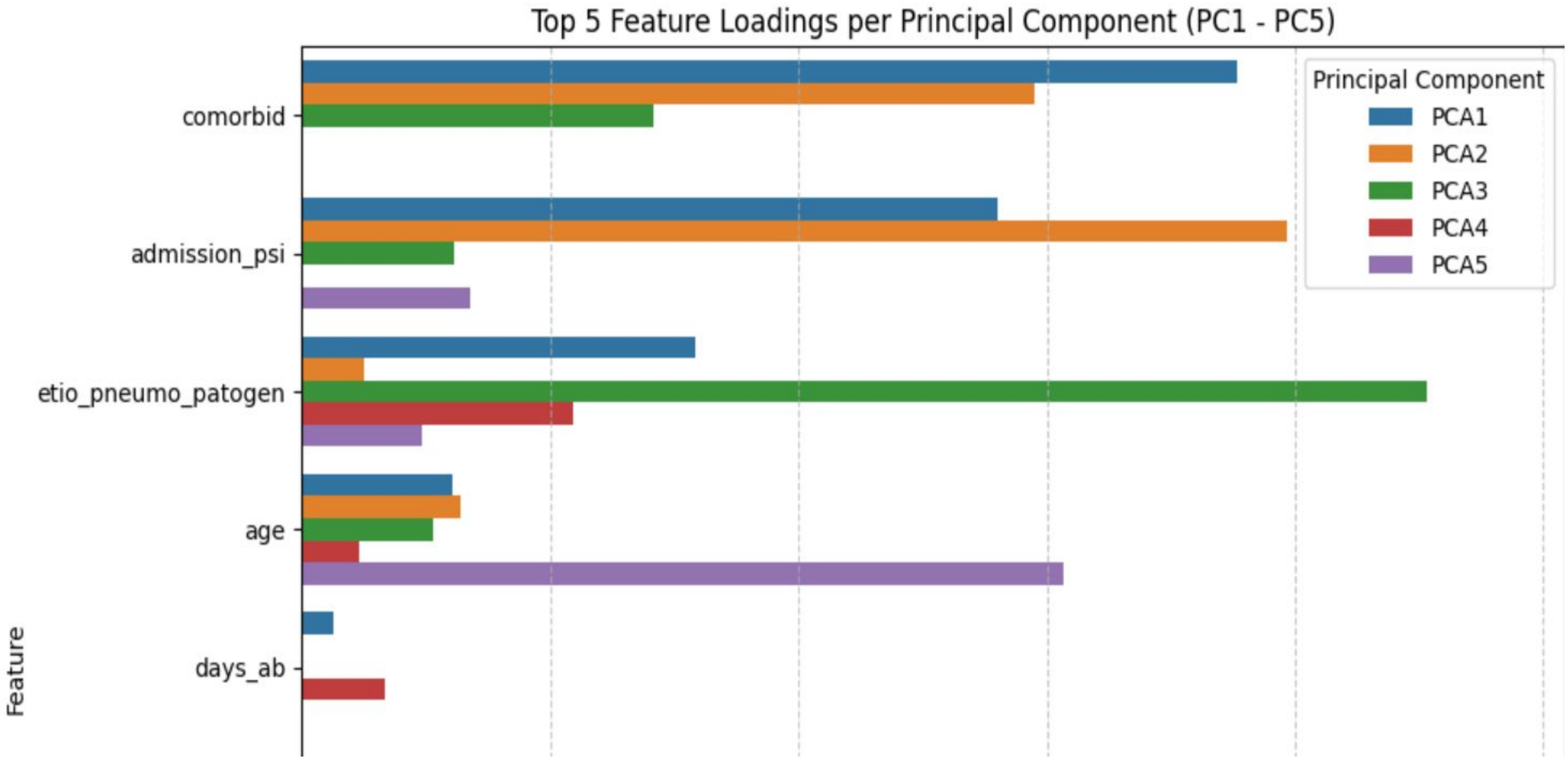
Principal Component	Top Influential Features
PC1	comorbid, admission_psi, etio_pneumo_patogen, age
PC2	admission_psi, comorbid, age, sofa_72
PC3	etio_pneumo_patogen, comorbid, dicharge_date
PC4	dicharge_date, etio_pneumo_patogen, days_ab, weight
PC5	age, weight, height, etio_pneumo_patogen

Performance Metrics

Model Performance Comparison

Model	Accuracy	Precision (Weighted Avg)	Recall (Weighted Avg)	F1-Score (Weighted Avg)
Logistic Regression	0.933	0.96	0.93	0.95
Random Forest	0.966	0.96	0.97	0.96
XGBoost	0.980	0.96	0.98	0.97

PCA components



PCA components

Principal Component	Top Influential Features	Interpretation
PC1	comorbid, admission_psi, etio_pneumo_patogen, age	Captures patient condition severity and comorbidity—likely distinguishing patients with complex health profiles. Comorbidity refers to the simultaneous presence of two or more medical conditions in a patient.
PC2	admission_psi, comorbid, age, sofa_72	Emhasizes clinical scores(sofa 72) and patient age—important for assessing initial health status.
PC3	etio_pneumo_patogen, comorbid, dicharge_date	Focuses on pneumonia pathogen types and comorbidity—relevant to diagnosis and outcome.
PC4	dicharge_date, etio_pneumo_patogen, days_ab, weight	Captures discharge-related info and antibiotic days—indicative of treatment duration or recovery time.
PC5	age, weight, height, etio_pneumo_patogen	Likely represents demographic and anthropometric variability.

Conclusion

- Models performed well though they significantly favoured the majority class even after applying SMOTE.
- This project was a gateway into advanced AI in Pneumonia research, it proved that clinical data can also be applied to detecting pneumonia to some extent.

Future Direction

- Combine Structured clinical data with image based data for predictive modeling.
- Apply deep learning techniques such as transfer learning to curb the issue of lack of data.
- Apply Federated Learning, to expose the model to diverse data for enhanced generalizability.
- Deployment. This involves scaling the project to a full-fledged application.

References

1. Effah, C. Y., Miao, R., Drokow, E. K., Agboyibor, C., Qiao, R., Wu, Y., Miao, L., & Wang, Y. (2022). Machine learning-assisted prediction of pneumonia based on non-invasive measures. Frontiers in Public Health, 10. https://pmc.ncbi.nlm.nih.gov/articles/PMC9371749/

2. Luo, Y., Tang, Z., Hu, X., Lu, S., Miao, B., Hong, S., Bai, H., Sun, C., Qiu, J., Liang, H., & Ning Na. (2020). Machine learning for the prediction of severe pneumonia during posttransplant hospitalization in recipients of a deceased-donor kidney transplant. Annals of Translational Medicine, 8(4), 82–82. https://doi.org/10.21037/atm.2020.01.09

3. Swetha, K. R., M, N., P, A. M., & M, M. Y. (2021). Prediction of Pneumonia Using Big Data, Deep Learning and Machine Learning Techniques. 2021 6th International Conference on Communication and Electronics Systems (ICCES). https://doi.org/10.1109/icces51350.2021.9489188

References

4. Jeon, E.-T., Lee, H. J., Park, T. Y., Jin, K. N., Ryu, B., Lee, H. W., & Kim, D. H. (2023). Machine learning-based prediction of in-ICU mortality in pneumonia patients. Scientific Reports, 13(1), 11527.https://www.nature.com/articles/s41598-023-38765-8

5. Bhattarai, P., Kumar, V., Th, B., & Rashmi Od. (2023). Detection of Pneumonia using Machine Learning. https://doi.org/10.1145/3647444.3652464

Thank you