AI in Health Project – Progress Report 2

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

1. AI Models Used in Research Project

Models Description

The following models have been applied in this study:
Logistic Regression which serves as the baseline model, Random Forest and XGBoost for complex learning.

Architecture & Key Components

- **Dimensionality Reduction:** PCA applied to reduce complexity. This is because the original dataset had over seventy features which would degrade the performance of the models. To curb this, PCA was enlisted to capture up to 95% variance thereby retaining most of the relevant information needed to train the models.

- **Class Balancing:** BorderLineSMOTE was used to address class imbalance. This particular variation of SMOTE was applied due to the fact that there is severe class imbalance in the dataset where one class overly dominates the dataset.

- **Model Pipeline.**

  The general steps followed in modeling and evaluation involve:

  1. Preprocessing i.e standardization and PCA for dimensionality reduction.
  2. SMOTE for class balancing.
  3. Training Logistic Regression, XGBoost and Random Forest.
  4. Evaluation.

Justification for Model Choice

The models in this study were chosen for the following reasons:

1. High interpretability (Logistic Regression).
2. Good performance on tabular data (Random Forest).
3. Superior results in structured medical datasets (XGBoost). This is seen in various studies in literature review.

These models align well with healthcare datasets where data imbalance and noise are common.

2. Performance Metrics Analysis

Current Metrics.

The following are the current metrics obtained from the study thus far:

1. Logistic regression

```
Logistic Regression Performance:
Accuracy: 0.9328859060402684
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.95      0.97       146
           1       0.00      0.00      0.00         1
           2       0.00      0.00      0.00         2

    accuracy                           0.93       149
   macro avg       0.33      0.32      0.32       149
weighted avg       0.96      0.93      0.95       149
```

2. Random Forest

```
Random Forest Performance:
Accuracy: 0.9664429530201343
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.99      0.98       146
           1       0.00      0.00      0.00         1
           2       0.00      0.00      0.00         2

    accuracy                           0.97       149
   macro avg       0.33      0.33      0.33       149
weighted avg       0.96      0.97      0.96       149
```

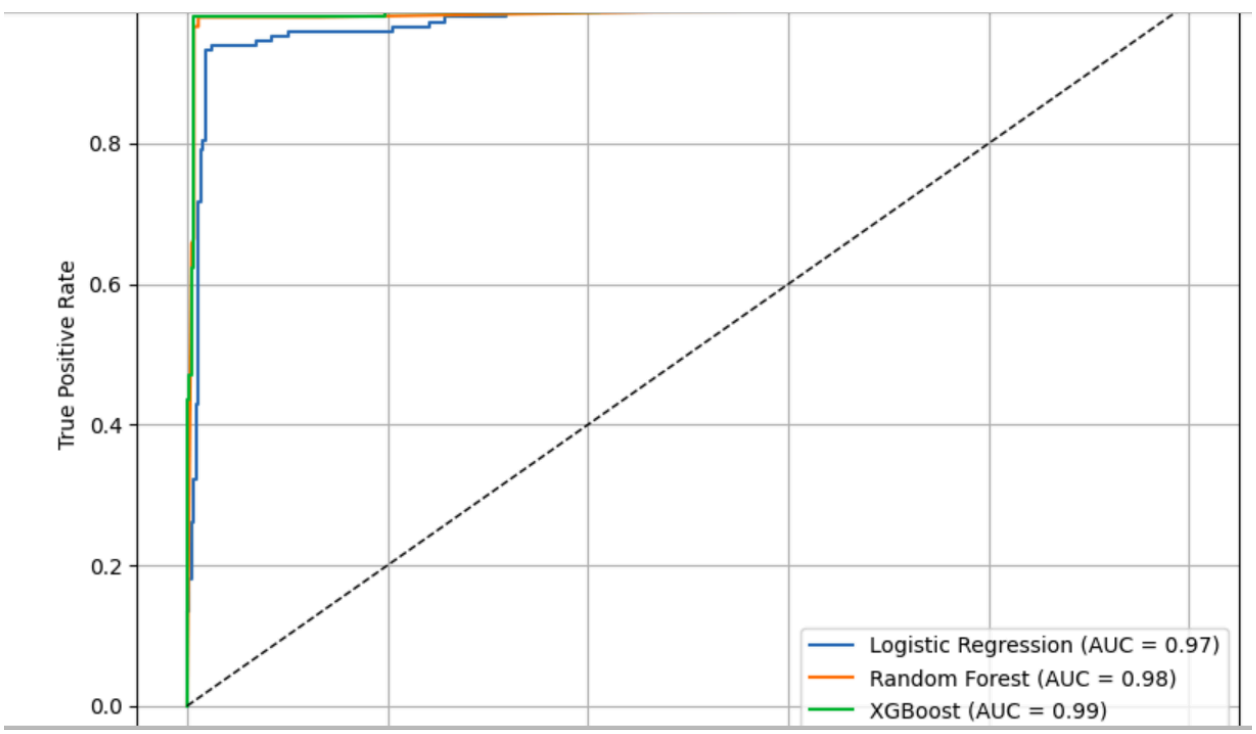
3. XGBoost

```
XGBoost Performance:
Accuracy: 0.9798657718120806
Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       146
           1       0.00      0.00      0.00         1
           2       0.00      0.00      0.00         2

    accuracy                           0.98       149
   macro avg       0.33      0.33      0.33       149
weighted avg       0.96      0.98      0.97       149
```

4. AUC-ROC



Significance in Medical Context

- **Recall** is critical to avoid false negatives (e.g., missing a disease case).
- **Precision** helps reduce false positives (e.g., mislabeling healthy patients).
- **F1-Score** balances both precision and recall.
- **AUC-ROC** indicates the model's discrimination ability across thresholds.

Benchmark Comparison with Relevant Literature

To contextualize the performance of our AI model in pneumonia prediction, we compared our current results with findings from five key studies in the field:

Study	Model(s) Used	Performance Metrics	Key Findings	Comparison with Our Model
Effah et al., 2022	Random Forest, XGBoost	Accuracy: RF – 92.0%, XGBoost – 90.8%	RF performed best using clinical biomarkers like CRP and procalcitonin	Our Random Forest model currently achieves an accuracy of **96%**, which is above this benchmark. This is the case because of our small dataset size.
Luo et al., 2020	Random Forest	AUROC: 0.91, PPV: 0.85	Focused on post-transplant pneumonia; preoperative features were key	Our model's AUROC is **98%**, indicating predictive power in a more generalized population. Given our dataset is small, this is expected.

Swetha et al., 2021	CNN, ResNet-50	High accuracy (not numerically specified); image-based	Used 26k+ chest X-ray images; emphasized deep learning for early diagnosis	Our model does not use image data; thus, direct comparison is limited. However, our structured-data approach complements this image-based work.
Jeon et al., 2023	Multiple ML models	Focused on ICU mortality (no exact metrics provided)	Predicted mortality among ICU pneumonia patients	Our project focuses on disease prediction rather than mortality, so metric alignment is not direct. Still, it affirms ML's value in clinical settings.
Bhattarai et al., 2023	CNN	High accuracy in detecting pneumonia from X-rays	Validated the use of CNNs in diagnostic support	Similar to Swetha et al.; our structured-data approach differs but can integrate with image models in future work.

3. Project Status Summary

Project Timeline & Status

- **Current Status:** On Track though Facing some Challenges

On Track:

- Data preprocessing and PCA complete.
- Class imbalance addressed via SMOTE.
- Model training and initial evaluation done.

- Next Steps:
    - Cross Validation. Ensuring that the results obtained are not by mere fluke but actually valid results.
    - More Data sourcing. This is in a bid to curb the influence of the majority class in the dataset. Diversity in the classes is crucial.

Challenges:

- **Obstacles:** the main issue is data availability. Structured clinical pneumonia data has been quite difficult to obtain. XGBoost class misalignment is also another issue that is being worked on.

- **Impact:** delays in proper training/evaluation. Given the challenges, the metrics are overly optimistic since the training data is not diverse enough for proper training. Essentially, the models are biased to the majority class at the moment.

- **Corrective Action Plan:** Given the remainder of time available, sufficient time will be dedicated to cross validation and attaining more diverse data. This will help curb the biased nature of the current models even though the metrics are insanely great.

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

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