Project Group 23

Daniel Akama Nyamweya

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

Introduction

Pneumonia remains a significant global health burden, responsible for millions of hospitalizations and deaths annually. Despite advances in medical imaging and diagnostic tools, early and accurate detection of pneumonia remains a challenge, particularly in resource-limited settings where access to radiologists and specialized equipment is scarce. Delayed or incorrect diagnosis can lead to severe complications, prolonged hospital stays, and increased mortality rates.

Machine learning (ML) has emerged as a powerful tool in medical diagnostics, offering the potential to enhance pneumonia detection through automated, data-driven models. By leveraging patient data, clinical biomarkers, and imaging techniques, ML algorithms can provide rapid and accurate predictions, aiding clinicians in decision-making. The integration of ML models in healthcare could reduce diagnostic errors, optimize treatment strategies, and improve patient outcomes.

This research aims to develop a robust ML model for pneumonia prediction using a structured dataset. Unlike conventional methods relying solely on chest X-ray interpretation, this study explores a broader set of clinical features to enhance prediction accuracy. By implementing and evaluating various ML algorithms, this study seeks to identify the most effective approach for early pneumonia detection, ultimately contributing to improved healthcare efficiency and patient care.

Literature Review

In this section we look at various papers on projects surrounding pneumonia prediction and their relevance to this research project. The following are summaries of the relevant literature:

1. **"Machine learning-assisted prediction of pneumonia based on non-invasive features"**. This study investigated the application of eight machine learning models to predict pneumonia using biomarkers, laboratory parameters, and physical features. The research involved 535 patients, each with 45 features. The Random Forest (RF) and XGBoost models demonstrated the highest performance, achieving accuracies of 92.0% and 90.8%, respectively. Key biomarkers such as C-reactive protein and procalcitonin were identified as significant predictors. The study concluded that ensemble machine learning models are effective in diagnosing pneumonia when individual clinical indicators are insufficient. (Effah et al., 2022)

2. **"Machine learning for the prediction of severe pneumonia during post-transplantation periods"** . This study focused on predicting severe pneumonia in patients post-kidney transplantation using machine learning algorithms. Out of 519 transplantations, 43 episodes of severe pneumonia were identified. The Random Forest classifier outperformed other models, showing an area under the receiver operating characteristic curve (AUROC) of 0.91 and a positive predictive value (PPV) of 0.85. Important predictive features included preoperative pulmonary lesions and reoperation. The study highlighted the potential of machine learning in enhancing predictive performance for severe pneumonia in post-transplant patients. (Luo et al., 2020)

3. **"Prediction of Pneumonia Using Big Data, Deep Learning and Machine Learning Techniques"**. This research developed a robust machine learning model for early pneumonia detection using chest X-rays. The study employed a convolutional neural network (CNN) model and a pretrained ResNet-50 model. The dataset comprised 26,684 chest X-ray images from unique patients. The models aimed to identify pneumonia patterns, enabling prompt diagnosis and treatment. The study emphasized the effectiveness of deep learning algorithms in medical image analysis for pneumonia detection. (Swetha et al., 2021)

4. **"Machine learning-based prediction of in-ICU mortality in pneumonia patients".** This study aimed to develop machine learning models for predicting mortality in patients with severe pneumonia admitted to intensive care units (ICUs). The researchers utilized various algorithms to analyze patient data and identify factors associated with increased mortality risk. The study demonstrated that machine learning models could effectively predict in-ICU mortality, aiding in clinical decision-making and resource allocation. (Jeon et al., 2023)

5. **"Detection of Pneumonia using Machine Learning"**. In this study, researchers proposed a system for automatically identifying pneumonia from chest X-ray images using deep learning algorithms. The approach involved training convolutional neural networks (CNNs) to detect pneumonia with high accuracy. The study highlighted the potential of deep learning techniques in improving diagnostic accuracy and assisting healthcare professionals in the timely detection of pneumonia. (Bhattarai et al., 2023)

The studies reinforced the decision to undertake this research project, mainly due to the fact that highly-dimensional structured clinical datasets have not really been applied in multi-class pneumonia detection. Application of machine learning in this area is promising.

Data Collection and Methodology

Dataset

Data collection is an ongoing process. It involved going through various data repositories on the internet to find structured pneumonia datasets. Most require permission to access and download.
Structured pneumonia datasets are limited on open repositories such as Kaggle since most datasets there are image based. We sourced structured data from sources such as MIMIC-III, NLM and Physionet. The dataset found came from Physionet, it had a total of 85 columns and 769 rows. It contains patient records with features such as: patients age, gender, BMI, whether they are a health worker, vaccination information, whether the patient has received antibiotics in the last 12 months, smoking, COVID diagnosis, patients cardiovascular health history etc.

The target variable i.e. main_diagnosis column is what the model will be predicting. The target is multi-class consisting of seven diagnosis of pneumonia e.g. Community-acquired pneumonia, ventilator-associated pneumonia, healthcare-associated pneumonia, Ventilator-Associated Tracheobronchitis (VAT), Non-ICU-acquired Pneumonia not associated with VM, Non-ICU-acquired Pneumonia not associated with VM, VM-associated Pneumonia not acquired in ICU. Link to dataset .

The dataset was deemed ‘good data’ using the following criteria:

1. Relevance to the Problem - Does the dataset contain the features needed for your pneumonia prediction task? The Dataset presents a huge challenge since most of the features are relevant to the study. Adequate steps will be taken to prevent the curse of dimensionality.

2. Data Quality & Completeness - checking for Missing Values and assessing if they can be reasonably imputed. Inconsistencies, errors like incorrect labels, duplicate records, or outliers.

3. Feature Selection & Importance - whether the included features are relevant to pneumonia diagnosis. Are there redundant or highly correlated features?

4. Ethical & Legal Considerations - Does the dataset comply with privacy regulations (e.g. HIPAA, GDPR) if it contains patient data? Is there potential bias in the data that could lead to unfair predictions (e.g. underrepresentation of certain populations) and what are the data sources? Is it from a reliable institution?

Methodology

This methodology ensures a structured approach that will be applied to building a reliable pneumonia model (future steps). The evaluation step plays a critical role in validating the effectiveness of the model.

1. Data Preprocessing

Before training the model, the dataset will undergo cleaning and transformation to ensure quality and consistency.

1.1 Handling Missing Data

- Remove records with excessive missing values.
- Impute missing values using mean/mode for numerical data and most frequent values for categorical data.

1.2 Feature Engineering

- Normalize numerical features (e.g., Min-Max Scaling or Standardization).
- Convert categorical variables (e.g., gender, smoking status) using one-hot encoding.
- Apply feature selection techniques (e.g., Mutual Information, Recursive Feature Elimination) to remove irrelevant features.

1.3 Addressing Class Imbalance

- Use **SMOTE (Synthetic Minority Over-sampling Technique)** or class weighting to balance pneumonia cases.

2. Model Development

Multiple machine learning algorithms are trained and compared to identify the best-performing model for pneumonia classification.

2.1 Selected Algorithms

- Logistic Regression (Baseline Model).
- Random Forest Classifier.
- XGBoost.

Random Forest and XGBoost models are to significantly improve prediction accuracy by combining the strengths of multiple individual models.

2.2 Model Training

- Train models using 80-20 train-test split.
- Use k-fold cross-validation (k=5 or 10) to ensure robustness.
- Optimize hyperparameters using Grid Search.

3. Model Evaluation

Performance evaluation is crucial to ensure the reliability of the model. The following metrics are used:

3.1 Classification Metrics

- **Accuracy:** Measures overall correctness of predictions.
- **Precision, Recall, and F1-Score:** To assess per-class performance.
- **AUC-ROC:** Evaluates how well each class is distinguished.

3.2 Benchmark Comparison

- Compare model performance against a simple baseline classifier (Logistic Regression).
- Assess results against clinical benchmarks or previous research studies.

3.3 Model Interpretability

- Use **SHAP (SHapley Additive Explanations)** or feature importance plots to explain model decisions.
- Identify key features influencing pneumonia classification.

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