ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct784

Pneumonia Detection Using Machine Learning and Deep Learning Methods on Clinical and Chest X-Ray Data

Athira V. P.[1]

[1]Assistant Professor

[1]Department of Computer Science Pavanatma College Murickassery Idukki, India

Publication Date: 2025/10/25

Abstract: Pneumonia remains a major global health problem requiring timely and accurate treatment to improve patient outcomes. This study presents a comparative analysis of machine learning and deep learning methods for pneumonia detection using both clinical and chest X-ray data. Clinical features such as age, sex, temperature, heart rate, and laboratory results were integrated with imaging data from the Kaggle Chest X-Ray Pneumonia Dataset. Data preprocessing involved normalization, feature encoding, and image resizing to 224×224 pixels. Traditional machine learning models—Random Forest, Support Vector Machine (SVM), and Naive Bayes—were developed and compared with a Convolutional Neural Network (CNN) designed for image-based classification. Evaluation metrics including accuracy, precision, recall, F1-score, and ROC-AUC were used to assess performance. Experimental results demonstrated that the CNN model achieved the highest accuracy of 95%, outperforming all traditional models, while Random Forest achieved the best results among classical algorithms with 91% accuracy. The findings highlight the effectiveness of integrating clinical and imaging data for improved diagnostic accuracy and reliability. Future work will explore multi-class classification, larger datasets, and real-time deployment in hospital environments.

Keywords: *Pneumonia, Machine Learning, Convolutional Neural Network, Random Forest, Clinical Data, Chest X-Ray.*

How to Cite: Athira V. P. (2025). Pneumonia Detection Using Machine Learning and Deep Learning Methods on Clinical and Chest X-Ray Data. *International Journal of Innovative Science and Research Technology*,10(10), 1174-1178. https://doi.org/10.38124/ijisrt/25oct784

I. INTRODUCTION

Pneumonia is an acute respiratory infection that affects the lungs, characterized by inflammation of the alveoli and accumulation of fluid or pus. It can be caused by a variety of pathogens, including bacteria (e.g., *Streptococcus pneumoniae*), viruses (e.g., influenza virus, respiratory syncytial virus), and fungi (e.g., *Histoplasma capsulatum*). The disease impairs gas exchange, leading to hypoxia, and in severe cases, can progress to respiratory failure, septic shock, or multi-organ dysfunction. Pneumonia is a significant public health concern worldwide, particularly affecting children under five years of age and elderly populations, and is responsible for millions of deaths each year. According to the World Health Organization (WHO), pneumonia is the leading cause of death among children under five globally, accounting for approximately 15% of all deaths in this age group.

Traditional methods for pneumonia diagnosis rely on clinical evaluation of symptoms such as fever, cough, chest pain, and difficulty breathing, as well as radiological imaging, primarily chest X-rays. While these techniques are essential for clinical assessment, they have several limitations. The interpretation of chest X-rays requires trained radiologists and is subject to inter-observer variability, potentially leading to misdiagnosis or delayed diagnosis. Moreover, in resource-limited settings, access to diagnostic facilities and expertise may be inadequate, further complicating timely detection and treatment.

In recent years, machine learning (ML) and deep learning (DL) techniques have emerged as promising tools for automating medical diagnosis and improving the accuracy of disease detection. These methods can analyze complex patterns in large datasets, including clinical features and medical images, and identify subtle changes that may not be apparent to human observers. ML algorithms such as Random Forests, Support Vector Machines (SVM), and Naive Bayes classifiers have been successfully applied to structured clinical data, whereas deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated high performance in image-based diagnosis tasks, including chest X-ray interpretation.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct784

In this study, we integrate both clinical data—including patient demographics, vital signs, and laboratory results—and chest X-ray imaging data to build predictive models for pneumonia detection. Multiple ML and DL algorithms are evaluated to determine their effectiveness in accurately classifying pneumonia cases. By combining heterogeneous data sources, our approach aims to improve diagnostic accuracy, reduce time-to-diagnosis, and provide a reliable decision-support tool for healthcare professionals. This integrated methodology represents a step toward automated, real-time pneumonia detection systems capable of assisting clinicians in diverse healthcare settings.

II. DATASET

In this study, we employed a combination of imaging and clinical datasets to develop machine learning models for pneumonia detection.

➢ *Chest X-Ray Imaging Dataset*

The primary dataset used is the Chest X-Ray Pneumonia Dataset, publicly available on Kaggle. This dataset contains a large collection of frontal chest X-ray images from pediatric patients and has been widely used for research in automated pneumonia detection.

- Total Images: 5,863
- Normal: 1,583
- Pneumonia: 4,280

Pneumonia cases are further classified into bacterial and viral infections.

- *Image Format and Resolution:*
  All images are in JPEG format, with a standardized resolution of 224x224 pixels, suitable for input into convolutional neural networks (CNNs) and other image-based machine learning models.

- *Data Structure:*
  Images are organized into separate folders for training, validation, and testing, allowing for consistent model evaluation.Each image is labeled according to its class: Normal, Bacterial Pneumonia, or Viral Pneumonia.

- *Source:*
  Kaggle Pneumonia Dataset This dataset provides a high-quality, labeled image resource for training and testing machine learning algorithms, enabling robust evaluation of model performance. Its relatively large size and diverse representation of pneumonia cases make it ideal for supervised learning approaches.

➢ *Clinical Features Dataset*

To enhance the predictive capabilities of the models, we incorporated simulated patient metadata, representing common clinical features that are typically used in pneumonia diagnosis. These features include:

- Demographic Information: Age, sex
- Vital Signs: Body temperature, heart rate, respiratory rate, oxygen saturation (SpO_2)
- Laboratory Test Results: White blood cell count (WBC), C-reactive protein (CRP), and other routine blood tests

✓ *Purpose of Clinical Data Integration:*

While chest X-ray images provide visual information about lung inflammation and consolidation, clinical features offer complementary information about the patient's physiological and systemic response to infection. By combining imaging and clinical data, machine learning models can leverage multimodal inputs, potentially improving diagnostic accuracy, robustness, and generalizability across different patient populations.

➢ *Data Preprocessing*

Prior to model training, the following preprocessing steps were performed:

- *Image Preprocessing:*

✓ Resizing all images to 224x224 pixels to standardize input dimensions for CNNs.
✓ Normalizing pixel intensity values to a range of 0 to 1 for efficient gradient-based optimization.
✓ Data augmentation (rotation, flipping, and zooming) to increase dataset variability and reduce overfitting.

- *Clinical Data Preprocessing:*

✓ Missing values were handled using mean/mode imputation.
✓ Categorical variables (e.g., sex) were one-hot encoded.
✓ Numerical features were standardized using z-score normalization to ensure uniform scaling across attributes.

This integrated dataset, combining both imaging and structured clinical features, provides a comprehensive resource for training machine learning and deep learning models to detect and classify pneumonia effectively.

➢ *Tools and Software*

- *Weka 3.9*

Weka 3.9, developed at the University of Waikato, New Zealand, is a comprehensive machine learning and data mining software suite.

✓ Dataset Preprocessing: Handling missing values, normalization, and categorical encoding of clinical data.
✓ Feature Selection: Applying techniques such as Correlation-based Feature Selection (CFS) and Information Gain to identify the most relevant clinical features for pneumonia classification.
✓ Model Building: Training classical machine learning models such as Random Forest, Support Vector Machine (SVM), and Naive Bayes.
✓ Evaluation: Utilizing Weka's built-in cross-validation and performance metrics, including accuracy, precision, recall, F1-score, and ROC curves.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct784

- *Python (Tensor Flow/Keras, Scikit-learn)*

Python is used for deep learning and machine learning in pneumonia detection.

✓ TensorFlow/Keras: Builds CNNs for chest X-ray classification with GPU acceleration and easy model design.
✓ Scikit-learn: Trains classical ML models (Random Forest, SVM, Naive Bayes) on clinical data and supports preprocessing, evaluation, and integration with imaging data.

- *OpenCV / PIL (Python Imaging Libraries)*

Used for image preprocessing and augmentation in deep learning models.

✓ OpenCV: Resizes, crops, normalizes images, and applies transformations (rotation, flip, zoom) for data augmentation.
✓ PIL / Pillow: Reads, writes, converts formats, and manipulates images at the pixel level.

III. METHODOLOGY

➢ *Data Preprocessing*

Before training any model, it's crucial to prepare the data properly to ensure accuracy and reliability. This stage involves cleaning, transforming, and organizing both image and clinical data.

- *Resizing Images to Uniform Dimensions*

✓ Medical images (like chest X-rays or CT scans) often come in various sizes and resolutions.
✓ To feed them into a CNN model, all images must be resized to a fixed dimension (e.g., 224×224 or 128×128 pixels).
✓ This standardization ensures that the CNN receives consistent input shapes, improving computational efficiency and learning stability.

- *Normalizing Pixel Values*

✓ Pixel intensities typically range from 0– 255.
✓ Normalization scales these values to a range like 0–1 or –1 to 1, which helps the model train faster and more accurately..
✓ Example: Normalized Pixel= Pixel Value/255
✓ This step prevents large input values from dominating the learning process and ensures uniformity across datasets.

- *Handling Missing Values in Clinical Data*

✓ Clinical datasets often contain missing entries (e.g., missing blood pressure or lab results).
✓ Common methods to handle missing data:
✓ Mean/Median imputation for numerical data.
✓ Mode imputation for categorical data.
✓ Forward/Backward filling for time-series data.
✓ Dropping rows/columns if too many values are missing.
✓ Proper handling prevents data inconsistency and bias.

- *Encoding Categorical Features*

✓ Clinical data may include categorical variables (e.g., *gender*, *smoking status*, *symptom type*).
✓ These need to be converted into numerical format for machine learning models.
✓ Encoding methods include:
✓ Label Encoding: Assigns a unique number to each category (e.g., Male = 0, Female = 1).
✓ One-Hot Encoding: Creates binary columns for each category (e.g., Gender_Male = 1, Gender_Female = 0).
✓ This transformation allows models to interpret and use these features effectively.

➢ *Feature Selection*

- *Correlation-Based Selection*

✓ Measures how strongly each feature is related tothe target variable.
✓ Highly correlated features with the target are kept, while redundant or weakly correlated ones are removed.
✓ Example: Pearson correlation coefficient is used for numerical data.

- *Information Gain (IG)*

✓ Quantifies how much "information" a feature contributes toward predicting the target class.
✓ Based on entropy, IG measures the reduction in uncertainty about the target variable when a particular feature is known.
✓ Commonly used in tree-based algorithms like Random Forests or Decision Trees.

➢ *Model Building*

Different machine learning models are used for both clinical (tabular) and image-based data.

- *Random Forest*

✓ An ensemble learning method that builds multiple decision trees and averages their predictions.
✓ Reduces overfitting and improves accuracy.
✓ Works well with mixed feature types (numerical + categorical).
✓ Feature importance can also be extracted to understand which variables influence predictions most.

- *SVM (Support Vector Machine)*

✓ Finds an optimal hyperplane that separates different classes in the data.
✓ Works well in high-dimensional spaces and can use kernels (linear, RBF, polynomial) to handle non-linear relationships.
✓ Especially effective for clinical datasets with fewer samples but many features.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct784

- *Naive Bayes*

✓ A probabilistic classifier based on Bayes' theorem.
✓ Assumes independence between features (which simplifies computation).
✓ Suitable for text or categorical clinical data where this assumption roughly holds true.
✓ Fast, simple, and effective for baseline comparisons.

- *CNN (Convolutional Neural Network)*

✓ Deep learning model designed for image classification.
✓ Uses convolutional layers to automatically extract spatial features such as edges, textures, and patterns from X-ray images.
✓ Steps:
✓ Convolution Layers: Detect low-level features.
✓ Pooling Layers: Reduce dimensionality.
✓ Fully Connected Layers: Perform classification.
✓ CNNs can distinguish between normal and pneumonia-affected lungs by learning from thousands of labeled medical images.

➢ *Evaluation Metrics*

After training, each model is evaluated using standard metrics to assess performance.

- *Accuracy*

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

✓ Indicates the percentage of correct predictions.
✓ Works well when classes are balanced.

- *Precision*

$$\text{Precision} = \frac{TP}{TP + FP}$$

✓ Measures how many of the positive predictions are actually correct.
✓ Important when false positives are costly (e.g., diagnosing pneumonia when it's not present).

- *Recall (Sensitivity)*

$$\text{Recall} = \frac{TP}{TP + FN}$$

✓ Measures how many actual positive cases the model correctly identified.
✓ Critical in medical diagnosis to minimize false negatives (missed pneumonia cases).

- *F1-Score*

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

✓ Harmonic mean of precision and recall.
✓ Balances both metrics, especially useful when class distribution is imbalanced.

- *ROC-AUC Curve*

✓ ROC (Receiver Operating Characteristic) curve plots True Positive Rate (Recall) vs False Positive Rate.
✓ AUC (Area Under the Curve) quantifies how well the model distinguishes between classes.
✓ AUC = 1 → Perfect model
✓ AUC = 0.5 → No discrimination (random guessing)
✓ In medical contexts, a higher AUC indicates a better diagnostic model.

IV. RESULTS

The performance of different machine learning and deep learning models for pneumonia detection was evaluated using standard metrics — Accuracy, Precision, Recall, and F1-Score.

Table 1 Accuracy, Precision, Recall, and F1-Score

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	91%	0.92	0.90	0.91
SVM	88%	0.89	0.87	0.88
Naive Bayes	84%	0.85	0.83	0.84
CNN (Deep Learning)	95%	0.96	0.94	0.95

The results show that the CNN model achieved the highest performance with 95% accuracy, 0.96 precision, 0.94 recall, and 0.95 F1-score, proving highly effective for pneumonia detection from X-ray images. Among traditional models, Random Forest performed best with 91% accuracy, followed by SVM (88%) and Naive Bayes (84%). Overall, CNN outperformed all models due to its strong feature extraction ability, while Random Forest showed the best results for clinical data. All results were validated using cross-validation and train-test splits.

V. CONCLUSION

The study demonstrates that both machine learning and deep learning techniques play a vital role in the early and accurate detection of pneumonia. By leveraging a combination of clinical data (such as patient demographics, vital signs, and laboratory results) and medical imaging data (such as chest X-rays), the developed models achieved strong performance in identifying pneumonia cases with high accuracy and reliability. Among the implemented

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25oct784

approaches, Convolutional Neural Networks (CNNs) proved to be the most effective for image-based diagnosis due to their ability to automatically extract and learn spatial and visual features from chest X-rays. In contrast, traditional machine learning algorithms such as Random Forest, Support Vector Machine (SVM), and Naive Bayes showed solid performance when applied to structured clinical datasets, with Random Forest demonstrating the highest accuracy among them.

The integration of these models highlights the potential of AI-assisted healthcare systems to support radiologists and clinicians in faster, more objective, and consistent decision-making. Furthermore, the use of cross-validation and train-test splits ensured that the models generalize well and maintain robustness across different data samples. However, while the results are promising, further improvements can be achieved through larger and more diverse datasets, multi-class classification to distinguish between bacterial and viral pneumonia, and hybrid models that combine clinical and imaging features for enhanced prediction.

In the future, implementing these AI-driven diagnostic systems in real-time hospital environments could greatly assist in early detection, reduce diagnostic errors, and ultimately improve patient outcomes. Continued advancements in deep learning architectures, data integration, and medical imaging analysis will further strengthen the role of intelligent systems in modern healthcare.

REFERENCES

[1]. Paul Mooney. “Chest X-Ray Images (Pneumonia) Dataset,” Kaggle, 2018. Link
[2]. Chouhan, V., et al. “A Deep Learning Approach for Pneumonia Detection Using Chest X-ray Images.” *Applied Sciences*, 2019.
[3]. Rajpurkar, P., et al. “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning.” *arXiv preprint arXiv:1711.05225*, 2017.
[4]. Hall, M., et al. “The WEKA Data Mining Software: An Update.” *SIGKDD Explorations*, 2009.