FALL - 2024

PREDICTING PNEUMONIA FROM CHEST X-RAYS USING ML MODELS

UNIVERSITY OF HOUSTON

INDE7397 25957 - Introduction to Data Mining

Group_9: Project Report

Done By:

1. **Bhanu Rama Ravi Teja Gonugunta, (**2349943)
2. **Harshini Kodali,** (2348586)
3. **Praneetha Raju Buddharaju,** (2350205)

Executive Summary

Pneumonia continues to be one of the major challenges to health globally, thus requiring a timely and precise diagnosis to reduce very serious complications and mortalities. The report covers some discussions related to detecting pneumonia using machine learning techniques through chest X-ray images. In the following study, both classical and advanced machine learning models such as Logistic Regression, Support Vector Classifier (SVC), k-Nearest Neighbors (KNN), Naive Bayes, Random Forest (RF), K-Means Clustering, and Artificial Neural Networks (ANN) will be evaluated regarding their diagnostic capability. Key performance metrics include accuracy, sensitivity, specificity, and F1-score.

It therefore follows that the pre-processing of data, feature engineering, and optimization of models emerge as important steps in ensuring reliability and robustness, with findings showcasing the promise of machine learning as a complementary diagnostic tool in healthcare. The methodology of the report adopted a structured machine learning pipeline for chest X-ray image classification. This pipeline starts off with the acquisition and preprocessing of data, which involves cleaning, resizing, normalization, and transformation. Data augmentation techniques, such as rotation, flipping, and scaling, enhance model robustness by introducing variability in the training dataset. Models are trained on a publicly available labeled dataset, with performance comparisons underscoring the strengths and limitations of each approach.

SVC achieves the highest accuracy at 96%, followed by Random Forest at 93% and Logistic Regression at 91%, while KNN, Naive Bayes, and K-Means Clustering demonstrate lower effectiveness.

The results indicated that the more advanced machine learning models, especially the SVC and Random Forest, can yield a high diagnostic performance that can detect pneumonia from chest X-rays reliably. These findings point to the possibility of embedding machine learning into conventional diagnostic work processes, ensuring scalability and efficiency. On the other hand, the research underlines the necessity of high-quality preprocessing and the proper selection of models in order to guarantee accuracy for diverse patient populations and at different levels of image quality.

Section: 1 Background – Introduction

Pneumonia is a serious respiratory disease due to infections that inflame the air sacs of one or both lungs. The triggers can include bacteria, viruses, or fungi; the condition is one of the main causes of morbidity and mortality worldwide, especially in children under five and older adults. Early detection and timely treatment are critical in reducing the risk of complications, but traditional diagnostic methods include chest X-rays and lab tests that can be time-consuming, expensive, and prone to human interpretation errors. Machine learning and deep learning technologies have opened up new avenues for automation and improvement in pneumonia detection, allowing for quicker and more accurate diagnoses.

The project deals with the application of machine learning models to detect pneumonia from medical imaging data, specifically chest X-rays. The system aims to identify patterns and anomalies indicative pneumonia by training algorithms on labeled datasets. This approach reduces not only the dependency on human expertise but also increases accessibility, especially in low-resource settings where healthcare infrastructure is limited. In particular, automated detection systems have the potential to revolutionize diagnostic workflows, improve patient outcomes, and contribute to the global effort against this life-threatening disease.

Section: 2

A. Motivation:

The main aim of this Pneumonia Detection Test Project is to alleviate the increasing load of pneumonia in global health, especially in the deprived and resource-constrained regions of the world. Early detection becomes vital for preventing severe complications and a reduction in mortality rates. However, conventional diagnostic methodologies depend on human skills, which may not be always available and/or precise, especially in rural areas. This project uses machine learning and artificial intelligence to create an efficient, reliable, and low-cost solution for pneumonia identification from medical imaging data, including chest X-rays, to provide timely diagnosis and treatment.

This project also aims to fill the gap between technology and health by empowering automated systems to support health professionals who are overwhelmed with work. It also enumerates scalability, whereby AI-based solutions can analyze a huge volume of data with high precision, thus reducing diagnosis errors and improving patient outcomes. By democratizing access to advanced diagnostic tools, this initiative holds the potential to make a significant impact in saving lives globally.

B. Existing Methods:

Pneumonia, one of the leading causes of morbidity and mortality worldwide, is a bacterial, viral, or fungal infection that inflames air sacs in the lungs, which can be filled with fluids. Due to this fact, its diagnosis needs to be very accurate and timely for its effective management, which conventionally has presented difficulties because diagnosis relies on physical examination and chest X-rays, usually interpreted by radiologists. Artificial Intelligence and Machine Learning have brought in new methods of supplementing the conventional diagnostic workflow with computational models analyzing chest X-rays for pneumonia with greater acumen and speed. ChestX-ray8, introduced by Wang et al. in 2017, marked a beginning toward large-scale datasets in medical imaging, providing a hospital-scale database and benchmarks for thorax disease classification and localization [1].

This dataset has been instrumental in the development of weakly supervised learning techniques and forms the basis upon which ML-driven pneumonia detection systems can be developed. Expanding further on this work, more recent literature has investigated domain adaptation techniques to realize better transfer performance of models between different datasets, as illustrated by Feng et al. (2022) [2]. These approaches underpin the role of both supervised and unsupervised learning methods in their potential to improve diagnostic accuracy in a variety of clinical settings. Innovative deep learning architecture has further driven the field, enabling advanced image analysis for disease detection. Joseph and Anitha (2024) compared hybrid AI-driven models, revealing that integrated frameworks combining convolutional neural networks (CNNs) and traditional ML techniques can outperform standalone models [3].

Similarly, El-Kenawy et al. (2021) emphasized efficient pneumonia detection strategies tailored for chest radiography, highlighting the importance of lightweight yet effective algorithms in resource-constrained environments [4]. On the other hand, contemporary works by Kant (2024), Tshwale et al. (2024) demonstrate pre-trained models like ResNet50 to be quite robust regarding their efficiency for the extraction of meaningful features in complex medical images [5,6]. While the overall or holistic approaches of detection with remedies using Verma et al. (2024), where AI systems are used into the broader application of health applications [7].

C. Proposed Solution:

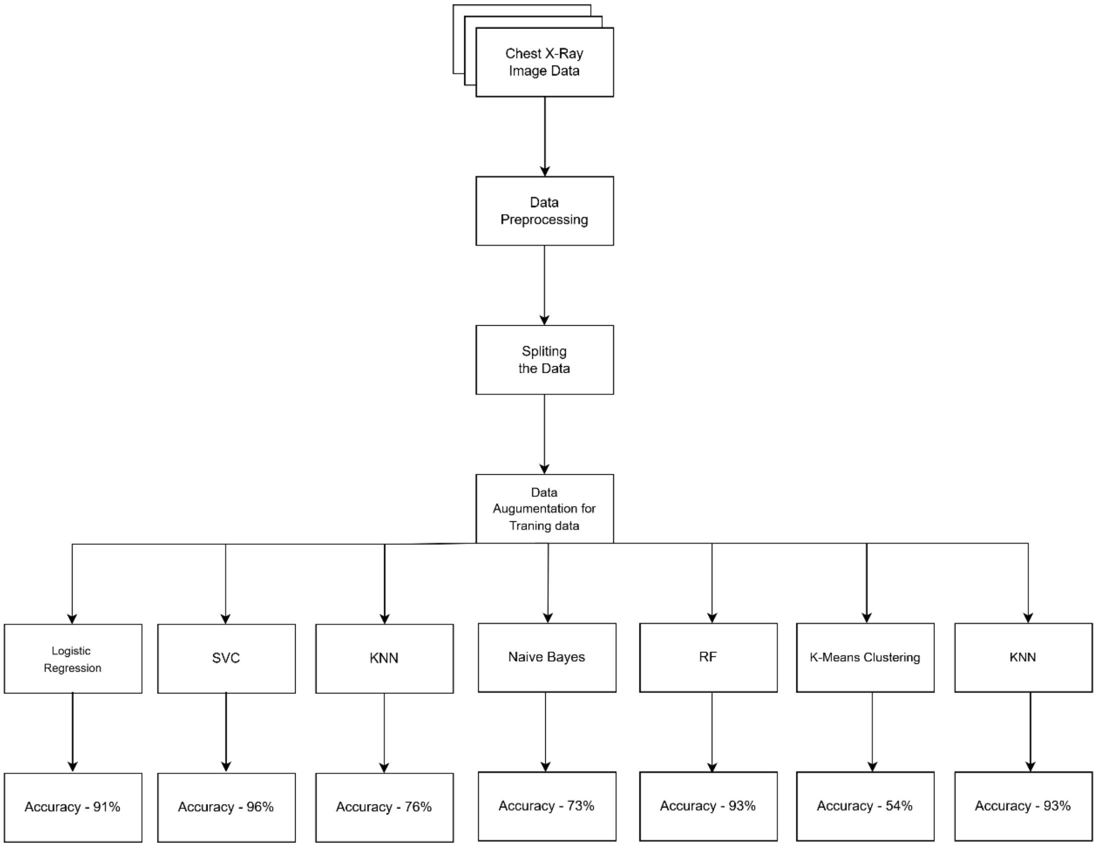

Fig.1 Proposed Model of Predicting Pneumonia

a. Chest X-Ray Data Set:

The Chest X-Ray dataset from Kaggle is a popular collection of medical imaging data used for diagnosing lung conditions like pneumonia, tuberculosis, or COVID-19. There are 5,863 X-Ray images. It contains 3 subfolders: Train, Test, and Validation, each containing grayscale X-ray images, with a total of labeled categories, including Normal and Pneumonia. These images are usually in JPEG format and vary in resolution; therefore, most of them require some preprocessing steps like resizing, normalization, or contrast enhancement to make them prepared for machine learning models. It supports classification, segmentation, and anomaly detection, among other tasks, and hence is very valuable for research and development in medical AI. Widely used examples include the Pneumonia Chest X-Ray dataset and the NIH Chest X-Ray dataset, which were shared for educational and research purposes in an ethical manner.

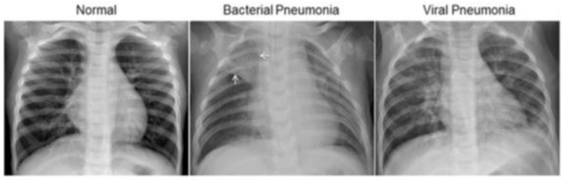

Fig.1 Examples of Chest X-Rays

Fig.1 shows Normal X-ray shows clear, dark lungs with no abnormalities. Bacterial Pneumonia presents localized, dense white opacities indicating infection in specific areas. Viral Pneumonia shows diffuse, patchy opacities spread across the lungs, affecting multiple areas uniformly.

b. Data Preprocessing:

- Loading and resizing images:

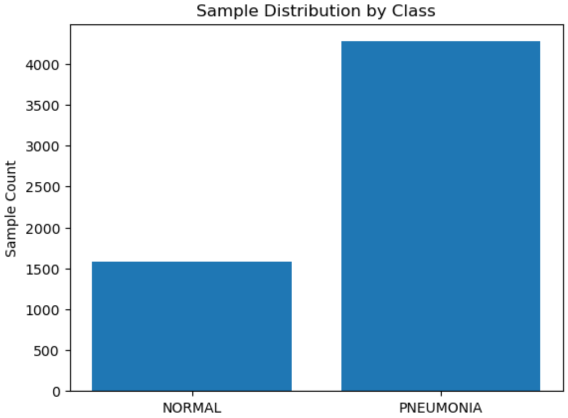

Fig.2 Sample Distribution

It reads images from labeled folders-NORMAL and PNEUMONIA-across test, train, and val directories; the code then numerically labels these images: 0 for NORMAL and 1 for PNEUMONIA. Preprocess the images by reading and turning them into grayscales, then resizing them to 200x200 pixels, adding to a dataset with their respective labels while skipping any missing or unreadable files. Next, shuffle the dataset to have random and unbiased training of a neural network. Compute the sample distribution on every class and plot it on a bar plot to indicate how many `NORMAL` versus `PNEUMONIA` samples one might get from the data as shown in Fig.2.

- Balancing the dataset:

Initially the data is in balance, now the data is converted to balances class distribution in the dataset using SMOTE of the imbalance package, which generates synthetic samples for the minority class. It flattens the images, applies SMOTE to perform oversampling, and reshapes data back into image form. It calculates the new class distribution and visualizes it as a bar plot as shown in Fig.3. The necessary packages include imblearn, numpy, matplotlib, and any preprocessing packages such as os and cv2.

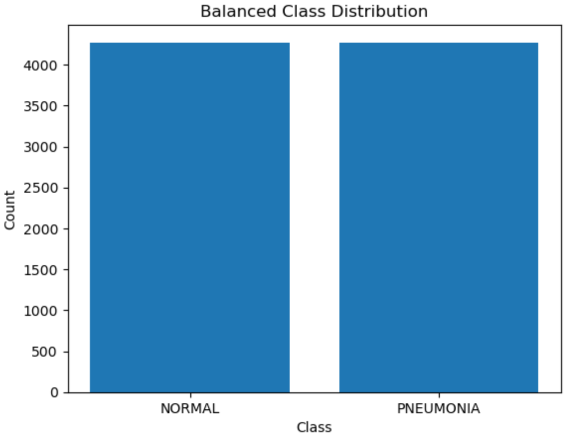

Fig.3 Sample Distribution

- Feature Scaling:

This code performs a check in the range of values in `x_train` before scaling to show the range in which initial data varies. After this, it proceeds with the work of Min-Max Scaling using `sklearn.preprocessing.MinMaxScaler`, which will normalize the data, therefore all values vary between 0 and 1. Then, to confirm a new range after scaling, it prints the maximum and minimum values for `x_train`. Scaling formula:

$$x_{Scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

c. Data Augmentation:

In our pneumonia detection project, we used the Chest X-ray or CXR dataset; it is a rich dataset with many labeled medical images. This data set consists of X-rays that are classified as either normal or with pneumonia from bacterial or viral infections. Diversity in patient demographics and various imaging conditions made this dataset very suitable for training and evaluating a machine learning model for the precise detection of pneumonia. However, with such a relatively limited dataset, overfitting could not be completely avoided, with poor model generalization, especially for the inherent variability in medical imaging.

The major strategy we employed for such challenges is data augmentation. Data augmentation artificially inflates the size of a dataset through various transformations, including rotations, flipping, zooming, change of brightness, and addition of noise. These techniques simulate real-world variations in imaging conditions and thus help the model become invariant with such changes. The introduction of diversity in the training data by augmentation not only enhanced the performance of the model but also improved its generalization ability for unseen test cases. This combination of the Chest X-ray dataset and augmentation formed the backbone of our approach, which helped in developing a reliable and effective pneumonia detection system.

d. ML Models:

In the Pneumonia Detection Test, different machine learning models were implemented, including Logistic Regression, SVC, KNN, Random Forest, Naive Bayes, ANN, and K-Means Clustering. These models were assessed using metrics such as accuracy, precision, recall, and F1 score to determine their effectiveness in identifying pneumonia from chest X-ray images. The SVC and Random Forest models showed the highest accuracy and reliability in pneumonia detection. This approach leverages AI to enhance diagnostic efficiency, reducing dependency on traditional methods and improving healthcare accessibility.

D. Experiments Settings and Comparison:

a. Confusion matrix and ROC Curve graph:

Fig. 4 to Fig. 10 shows the confusion matrix and ROC curve represent the performance of different models tested for the detection of pneumonia, which includes Logistic Regression, SVC, KNN, Naive Bayes, Random Forest, K-Means Clustering, and ANN. In the confusion matrix, classification accuracy for each model is represented through true positives, true negatives, false positives, and false negatives.

The ROC curve plots the true positive rate against the false positive rate, and the area under the curve, AUC, gives a measure of the general performance of the model. These plots give a complete comparison of the predictive capabilities of the models.

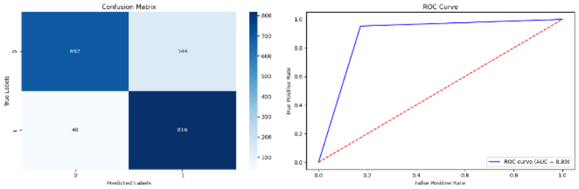

Fig.4 Confusion Matrix and ROC Curve for LR

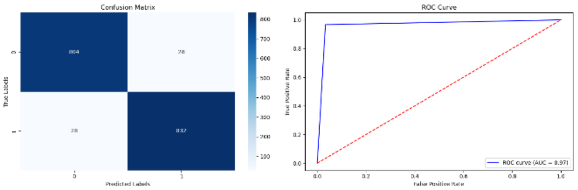

Fig.5 Confusion Matrix and ROC Curve for SVC

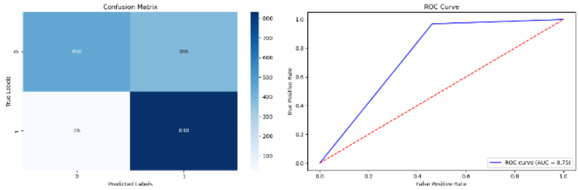

Fig.6 Confusion Matrix and ROC Curve for KNN

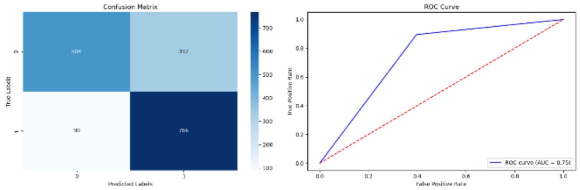

Fig.7 Confusion Matrix and ROC Curve for Navie Bayes

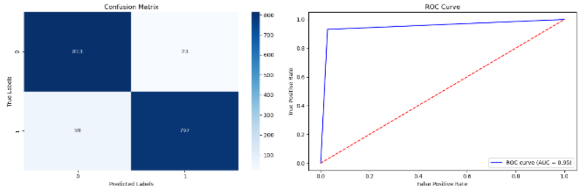

Fig.8 Confusion Matrix and ROC Curve for RF

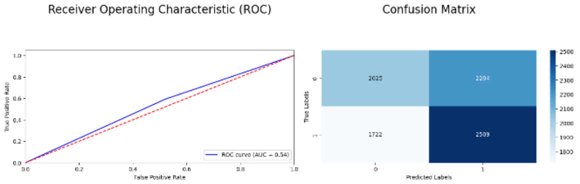

Fig.9 Confusion Matrix and ROC Curve for K-Means Clustering

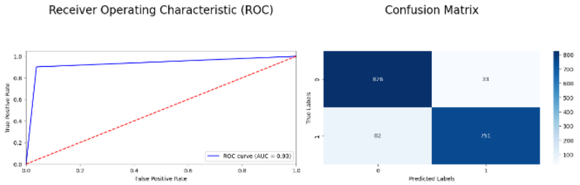

Fig.10 Confusion Matrix and ROC Curve for ANN

b. *Evaluation Matix and Comparison:*

Table I and Fig.11 are a detailed overview of comparisons among different machine learning models according to their accuracy A, precision P, recall R, and F1 score F1. By looking at the table below, the Support Vector Classifier (SVC) shows the top performance among other models on accuracy and an F1 score of 97 and 0.96, correspondingly. It is outperformed by Random Forest at 95% for the same, with an F1 score of 0.95. ANNs also show a high score of 93% in terms of accuracy and an F1 score of 0.93. On the contrary, K-Means presents the weakest performance, scoring just 54% accuracy. From this chart, the supremacy of the scores of SVC, RF, and ANN is easily underlined, as is the relatively poor performance presented by the K-Means clustering algorithm. These insights may indicate that for this dataset, SVC and RF are the most reliable models.

Table I. Results/Evaluation matrix of Pneumonia Detection

Model	A	P	R	F1
Logistic Regression	91%	0.96	0.86	0.91
SVC	**97%**	**0.95**	**0.95**	**0.96**
KNN	76%	0.82	0.77	0.75
Navie Bayes	73%	0.75	0.73	0.72
RF	95%	0.95	0.95	0.95
K-Means Clustering	54%	0.54	0.48	0.51
ANN	93%	0.93	0.96	0.93

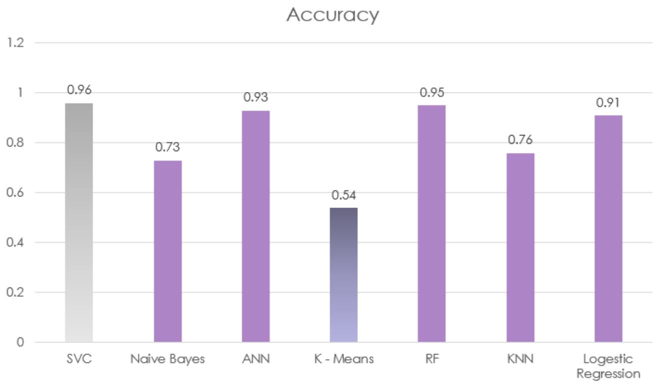

Fig.11 Comparison of ML Models using Bar Graph

Section: 3 - Summary

A. Knowledge Discovery:

Various machine learning models have been proposed for the detection of pneumonia using chest X-ray images. Logistic Regression, SVC, KNN, Random Forest, Naive Bayes, ANN, and K-Means Clustering were checked in terms of accuracy, precision, recall, and F1 score. The SVC proved to be the best among them all with an accuracy of 97% and a 0.96 F1 score, closely tagged by Random Forest and ANN. Confusion matrices and ROC curves were used for each model to obtain further insights into the performance of the model and the capability of models in distinguishing between positive versus negative cases. These findings emphasize the role of supervised learning techniques in medical imaging diagnostics.

B. Insights:

Performance analysis indicated that complex datasets could be handled using models like SVC and Random Forest. The performance of ANN also was exceptionally good, since it can capture intricate patterns in the medical images. But simple models such as KNN and Naive Bayes have poor results to achieve high precision and recall due to feature scaling sensitivity and imbalance in datasets. This may reflect the limited diagnostic utility of unsupervised approaches like K-Means Clustering in this context, while pointing out the importance of supervised methods. The ROC curves thus confirmed that both SVC and Random Forest models are robust and applicable for practical purposes.

C. Conclusion:

The study demonstrates the potential of machine learning algorithms, particularly SVC, Random Forest, and ANN, in improving the accuracy and efficiency of pneumonia diagnosis. This project develops a structured approach to testing and deploying AI-driven solutions in healthcare by integrating the powers of advanced models with visualization tools such as confusion matrices and ROC graphs. Future work may consider refining the models by having larger datasets, incorporating techniques to make AI explainable, or enhancing the usability of applications to low-resource settings as necessary for global impact.

REFERENCES

[1] Wang, X., Peng, Y., Lu, L., et al. "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." *IEEE CVPR*. 2017.

[2] Y. Feng et al., "Deep Supervised Domain Adaptation for Pneumonia Diagnosis From Chest X-Ray Images," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 3, pp. 1080-1090, March 2022, doi: 10.1109/JBHI.2021.3100119.

[3] H. Joseph and J. Anitha, "Enhanced Pneumonia Detection Through Advanced AI-Driven Hybrid Models: A Comparative Study of Deep Learning Architectures," 2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST), Kochi, India, 2024, pp. 01-05, doi: 10.1109/ICTEST60614.2024.10576088.

[4] El-Kenawy et al., Efficient Pneumonia Detection for Chest Radiography, International Journal of Intelligent Engineering and Systems, Vol. 14, No. 5, 2021.

[5] V. Kant, "A Comprehensive CNN-based Approach to Pneumonia Detection," 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2024, pp. 958-963, doi: 10.1109/ICOSEC61587.2024.10722467.

[6] M. Tshwale, P. Owolawi, M. Olaifa and V. Malele, "ResNet50 Pretrained Model Based Pneumonia Detection System," 2024 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 2024, pp. 0231-0236, doi: 10.1109/AIIoT61789.2024.10579006.

[7] K. Verma, P. Praveen and J. G. Ponsam, "Machine Learning for Holistic Pneumonia Management: Detection And Remediation," 2024 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES), Tumakuru, India, 2024, pp. 1-6, doi: 10.1109/ICSSES62373.2024.10561453.