Scientia Africana, Vol. 22 (No. 1), April, 2023. Pp 97-114 https://dx.doi.org/10.4314/sa.v22i1.10
© Faculty of Science, University of Port Harcourt, Printed in Nigeria ISSN 1118 – 1931

PNEUMONIA DISEASE DETECTION AND CLASSIFICATION SYSTEM USING NAIVE BAYESIAN TECHNIQUE

[1] Ojetunmibi, T., [2]Asagba, P. O., and [3]Okengwu, U. A.
[1,2,3]Department of Computer science, University of Portharcourt, Choba, Rivers State

Received: 24-03-2023
Accepted: 15-04-2023

ABSTRACT

Pneumonia is a chronic inflammation illness that affects both children and adults and is spread by various bacteria, viruses, and fungi. Since there are not enough specialists and facilities to interpret the findings of lab-based diagnosis, resulting to several cases of Pneumonia-related deaths. When the disease is discovered at an early stage as opposed to a later stage, it can be easily managed or controlled. The aim of the study is to create an effective pneumonia disease detection and classification system that uses Naive Bayesian and random forest Algorithms. The hash-based function was applied to train the model on X-ray chest samples from patients with pneumonia in order to improve detection accuracy and decrease classification errors. The hashing-based function was employed to compute and convert X-ray image features to a corresponding numerical code or label stored in a relative address and used as an array of reference given the associated values. The system was implemented using a future scaling technique that required the use of a hash encoding algorithm for the categorical labels of the target variable, and it improved model performance. We validated and compared the techniques in terms of accuracy and RMSE across different fine-tuned hyper-parameter values. The RF produced 97% with 3.33 error rate while NB recorded 99.08% accuracy rate as the best with 0.020 RMSE value.

Keywords: Feature selection,Naïve Bayessian (NB), Random forest (RF), Pneumonia disease

INTRODUCTION

Pneumonia is an autoimmune disorder brought on by a virus and fungus that mainly affects the bronchioles in the respiratory tract of human (Mcluckie 2009). Symptoms usually include a perfect blend of a dry or productive cough, chest pain, fever, and shortness of breath (Ashy and Turkington 2007). Infection with the respiratory failure disease has been documented throughout history of mankind (Raj and Prasanna 2012). However, pneumonia has been linked to millions of instances of pneumonia disease incidence rate and deaths globally, attracting the interest and attention of many medical professionals throughout the worldwide context today (Harshvardhan *et al.,* 2021). Rajpurkar et al. (2017) proposed a Radiologist-level pneumonia detection on chest x-rays with deep learning in order to detect pneumonia using x-ray images and a heat-map displaying the regions of FP, FN, TP, and TN cases detected for pneumonia disease (Liang and Zheng 2020). The pneumonia and how general public as a whole has dealt with the management and treatment of the disease globally may help us improve access to and the effectiveness of available treatment options and, in the long run, significantly lessen the negative effects that the illness causes (Chattopadhyay *et al.*, 2022). Annually, pneumonia claims the lives of over 1.5 million individuals globally, mostly in developing nations (Mabrouk et al, 20222). Numerous instances of Pneumonia disease-related deaths in both children and adults are caused by various bacteria, viruses, and fungi (Stephen *et al.,* 2019). Pneumonia is a disease that affects the majority of African communities, nations, and the entire world (Alsharif *et al.,* 2021). This is because there are not enough specialists (medical doctors) to interpret the results of a lab scientist's diagnosis using an x-ray, blood sample, and sputum culture. The easiest and most common imaging method for determining whether a patient has pneumonia is a chest X-ray. However, even for seasoned radiologists, diagnosing pneumonia and deciphering the results of chest radiography scan tests can be difficult, necessitating the use of a several machine learning techniques and algorithms. The low detection rate, noise, use of a complex model, and fluctuating patterns learned by the selected model limited the generalization ability of existing strategies with new Pneumonia Chest X-ray image dataset.The aim of this study is to design an efficient Pneumonia disease detection and classification system using naive Bayesian and random forest techniques. We intend to build, train, and test the Naive Bayes and RF classification models, then evaluate how well they perform at spotting infectious diseases features using the Chest X-ray dataset. The specific characteristics of the chest X-ray image data for pneumonia were computed and converted using the storage mapping hash function into a numerical code stored in a relative address in the hash table and used as an array of reference. The proposed system's hash-based index function, which maps Chest X-ray image features to a storage of key-values, will improve the model's detection accuracy and address issues with noise and outliers that limit existing approaches. The original CXR feature spaces in the learned hash code spaces can help detect disease features more precisely. The model will lessen the stress experienced by patients who have to search for and wait a long time for doctors to identify and interpret results of a pneumonia x-ray test obtained from clinical and pathological lab images. The use of the NB and RF machine learning model in the classification of the disease from the Pneumonia CXR Images as a novel innovation is due to its promising performance in both accuracy and efficacy.

The paper is organized as follows: Section 1 provides an introduction; Section 2 offers a brief assessment of prior approaches related to the topic and the gap in studying the proposed model; Section 3 introduces the model's materials and methods; Section 4 covers the results and a thorough discussion of the results; and Section 5 provides the paper's conclusion.

LITERATURE REVIEW

There are numerous related works that use machine learning knowledge and experience in Pneumonia disease classification and detection. Masud et al. (2021) created a novel methodology as the random forest (RF) classisification model to classify the Pneumonia disease using samples of chest x-ray images obtained from lab scientists.The RF technique's pre-processing stage was designed to extract statistical and global features from x-ray images. And in order to predict the target variable, which has three output variables, two very different features were combined using an RF model. To accurately measure performance, the suggested image dataset was rescaled with labels. The model was able to identify the type of pneumonia disease. The RF model generated an F1-Score value of up to 86.03% and accuracy metrics of 86.30% for classification.but was unable to distinguish between measurements of viral and bacterial pneumonia. Gupta (2021) suggested the

Scientia Africana, Vol. 22 (No. 1), April, 2023. Pp 97-114
© Faculty of Science, University of Port Harcourt, Printed in Nigeria
https://dx.doi.org/10.4314/sa.v22i1.10
ISSN 1118 – 1931

application of the convolution neural network (CNN) algorithm and transfer learning as a deep machine learning technique to analyze and detect Pneumonia disease using classification of chest x-ray images. In order to attain a high accuracy level, the dataset's size was increased using a technique called rotation and cropping. The CNN was adopted as a supervised classifier for detecting normal and abnormal scanned x-ray images. It was created to extract image features. Through computer vision, the CNN technique was able to accurately predict whether a patient has pneumonia or not with a validation accuracy of 93%. Comparatively speaking, the CNN method outperformed the transfer learning method. Ibrahim et al. (2017) developed a gradient boost algorithm-based LSTM auto-encoder technique to predict static feature. The model was employed to assess health-related risk and predict the occurrence of pneumonia. Performance was above average with a distribution accuracy of 89%, a 95% prediction accuracy, and 0.891 values for area under the curve. The fatality rate of the samples was 18.1%, with a range of 10.88% to 26.23%. The mortality rate was incredibly low (1.2 percent of cases), and there was no information provided about the model's recall parameter. It was not possible to compute the model recall parameter. Prayogo *et al.*, (2020) presented a siamese convolutional network (SCN) to build a better x-ray Pneumonia image classification model that can detect Pneumonia disease into the class of bacterial, normal and viral types. The SCN developed was trained to learn features about pairs of input images and produced 80.03% and 79.59% metrics of accuracy and f1-score value. The model required more training dataset to perform well and there is need to optimize image comparison to have higher accuracy rate. Kareem *et al.* (2022) aggregated two distinct machine learning techniques, such as KNN and CNN in the interest of identifying Pneumonia disease through the use of an image dataset. Transfer learning was utilized in conjunction with federated knowledge to aid experts in performing a consolidated approach for detecting Pneumonia disease using real-time datasets. A pre-trained deep CNN algorithm was proposed by Varshni *et al.*, (2019) to categorize and recognize normal and abnormal X-ray behaviours of Pneumonia disease emergence using image dataset. This was done before CNN model produced favorable results and a high rate of classification accuracies. A study was conducted by Katoch et al. in 2021 to deal with the different issues with the classification and detection of the pneumonia disease, which calls for proficiency in deep machine learning. The prominent risk variables for death rates around the world were predicted using the CNN, residual network (ReNet50), and a convolutional neural network with 16-layers called visual geometry group of 16-layers (VGG16) pre-trained models. **A** fully automated feature extraction from the X-ray images was accomplished by Alshehri *et al.*, 2021, using a framework for machine learning that included KNN, SVM, LR, NB, CNN, VGG16, and VGG19.This is done to evaluate the outlined Pneumonia disease's predictability and efficacy in the created model. The hybrid model was compared to the estimated adult infection rate for Pneumonia disease. Al-Dulaimi *et al.*, (2022) created a framework based on convolutional neural networks in order to detect Pneumonia disease using X-ray chest images. Accordingly, the precision, recall, f1-score, and detection accuracy rates were 98%, 98%, 97%, and 99.82%, respectively. The best accuracy for operations of diagnosis and improvement in disease management was found in the CNN-based Pneumonia detection model. Guleria and Sharma, (2023) merged VGG16 with SVM, RF and KNN, AIVE Bayesian and ANN for the purpose of identifying Pneumonia disease in elderly hospital patients with the same Chest-X-ray image dataset. The findings demonstrate that the proposed image dataset's ANN (VGG16) model outperformed the SVM (VGG16) model in terms of detection accuracy, with 92.15%, precision 0.9428, recall (0.9308), and f1-score (0.937). Abdullah *et al.*, (2022) developed and trained SVM with PCA and GA using 2,500-X-ray images for Pneumonia disease image classification in order to achieve high detection accuracy. X-ray images were processed using various methods of image analysis to extract the textual features, these were texture and shape-based features. El-Asnaoui, (2021) conducted investigations to build a multi-classification and Pneumonia disease detection system using inception ResNet version-2 and MobileNet version-2 with the aid of some self-study questionnaires. The Chest X-ray Image features were flattened to produce a fully interconnected neural network layer that can be categorized into the appropriate subcategory.

The whole section concentrates on the techniques used to identify Pneumonia disease from a dataset of x-ray images. The hash function is used in this study to extract features from Pneumonia image data at the pre-processing stage along with the pre-trained NB and RF classifiers. The proposed NB and RF implementation is broken down into a number of stages, including data collection, preprocessing, feature extraction, model building, training/testing, Pneumonia detection and evaluation. The NB and RF strategies are used along with a hash function encoder to minimize the computational task of categorizing features of the pneumonia disease from X-ray images into the normal or abnormal class,

Data collection: The dataset for this program was obtained from the "https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia" website of the UCI machine learning repository, which contains 2400 X-ray images. The Pneumonia image categorical variables of Finding_Labels, Single_Finding, Image_Index, validated, age_group, Patient_Gender and View_Position are encoded to numerical code using hashing based category. variables were transformed into distinct numerical codes.

MATERIALS AND METHODS

Table 1: Dataset (**Source:**https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia)

	Image_Index	Finding_Labels	- - -	Age_Group	Target
0	00009745_000.png	Cardiomegaly	- - -	(54.0, 64.0]	1
1	00015770_011.png	Cardiomegaly	- - -	(64.0, 91.0]	1
2	00015400_001.png	Cardiomegaly	- - -	(32.8, 43.0]	1
3	00011018_001.png	Cardiomegaly	- - -	(43.0, 54.0]	1
4	00000211_022.pn	Cardiomegaly	- - -	(54.0, 64.0]	1
- - -	- - - - - - --	- - - - - --	- - -	- - - - -	- - - -

2395	00009166_003.png	Pneumothorax	- - -	(52.0, 61.0]	1
2396	00017714_017.png	No Finding	- - -	(0.999, 29.0]	0
2397	00024435_000.png	No Finding	- - -	(40.0, 51.6]	0
2398	00021047_018.png	No Finding	- - -	(51.6, 61.0]	0
2399	00029880_008.png	Pneumothorax	- - -	(61.0, 85.0]	1

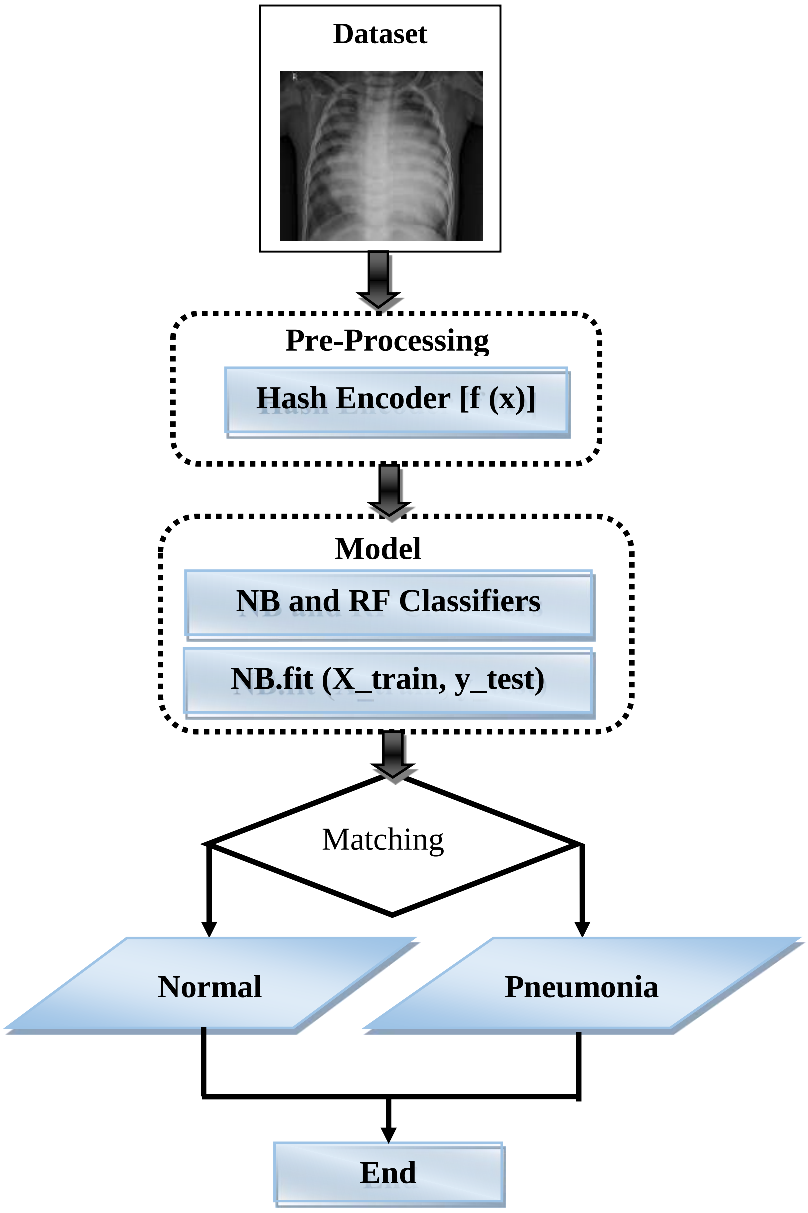


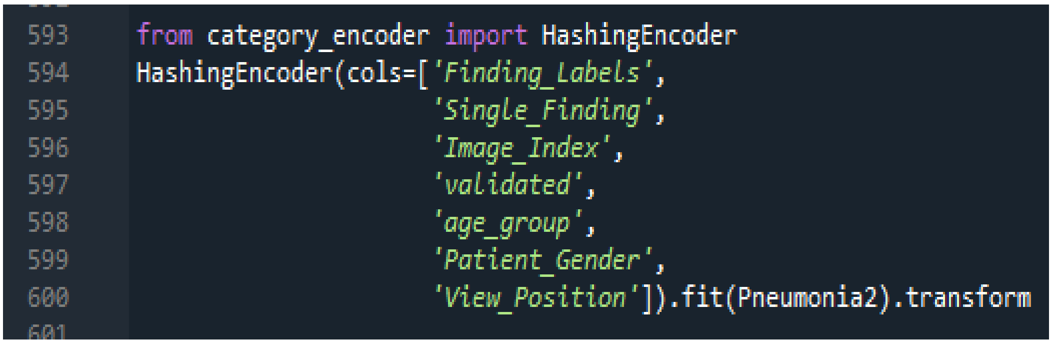
Figure 1 Study strategy (Masud *et al.*, 2021)

Preprocessing: Pre-processing is indeed the conversion of the dataset into a spotless dataset prior to feeding it into a machine learning algorithm and data collected from the outside world is frequently illogical, devoid of behavioral patterns, and may even be inaccurate. It involves the formatting of the data into a machine-understandable format through feature extraction, label encoding, and classification. Prior to the model development stages of training and testing, the raw data must be processed. It needs to be processed so that the system can understand it and the steps involved.

Feature extraction: Any application that relies on image preprocessing must use the feature extraction technique. This is the process of extracting and creating features to aid in the classification of objects. The feature vector is denoted as X where $X = (f_1, f_2,, ..., f_d)$ where f denotes features and d is the number of features extracted from character or digit based on the comparison of feature vector characters as been efficiently classified into appropriate class and recognized.

Classification is a task that calls for the application of machine learning (ML) methods that can identify classes based on samples of the problem domain. The classification scheme is used as a supervised learning method for determining the target or categories for data classes. A predictive task or modeling known as classification involves estimating a mapping function from discrete input variables (represented by "X" variables) to discrete output variables (represented by "y" variables). It primarily depends on the application area and the type of dataset that is available (Panjasuchat and Limpiyakorn, 2020). The model employs training data to convert the input data into class labels and numerical values and this was done using the feature hashing encoder function.

The Feature Hashing Encoder: We used a fixed size array and feature hashing to represent the dataset in a high-dimensional space. This is accomplished by using the hash function to encode the categorical feature data points of the Pneumonia x-ray Image dataset. The following segment of Python code as shown in Figure 2 was used to accomplish this.



```
from category_encoder import HashingEncoder
HashingEncoder(cols=['Finding_Labels',
                     'Single_Finding',
                     'Image_Index',
                     'validated',
                     'age_group',
                     'Patient_Gender',
                     'View_Position']).fit(Pneumonia2).transform
```

Figure 2: Hash function encoding

Random Forest (RF) Algorithm is an ensemble technique that creates a set of amazingly random classification or regression trees from a group of arbitrarily selected training data samples (Mahapatra (2014). The forest is created during the production process of trees (Abdulkareem and Abdulazeez, 2021). RF can be used to predict normal and Pneumonia disease affected patients in the classification model. The random forest classifier was called using the Python sklearn. ensemble library. The model was then trained using the training data, and the testing data was used to predict outcomes using NB. predict (X

Scientia Africana, Vol. 22 (No. 1), April, 2023. Pp 97-114 https://dx.doi.org/10.4314/sa.v22i1.10
© Faculty of Science, University of Port Harcourt, Printed in Nigeria ISSN 1118 – 1931

test). To determine whether hospital patients had the identified Pneumonia disease, the RF classifier was developed. In the Random Forest Classifier, the hyper-parameter values are adjusted or fine-tuned, and the number of estimators used to train the classifier ranges from 10, 20, 30, 40, to 70 random trees.

The Naive Bayes Model: The Gaussian Naive Probabilistic approach is used to compute the likelihood that data belong to a class given some prior knowledge. Given the class value, the probability calculations for each class are assumed to be conditionally independent. One of the most well-known supervised machine learning algorithms that makes use of the Bayes theorem is the Naive Bayes (NB) method. The Bayes theorem's definition of conditional probability serves as the foundation for the Gaussian NB classification algorithm. The conditional probability of an event "H" is given by the Bayes theorem, depending on whether event "D" has already happened. Based on prior knowledge of circumstances that might be related to the event, the Bayesian theorem essentially calculates the conditional probability of the occurrence of an event (Ramakrishnan *et al.*, 2018). We created and tested the NB model for accuracy in predicting the target variable on the basis of the training dataset. The sklearn. Naïve bayeslibrary in Python was used to call the classifier and create it. The training dataset was used to train the model, and the testing dataset was used to predict results NB. predict (X test). The mat plot lib library for Python was used to create the visualization. With the pre-processed training data, a Gaussian NB classifier was built to determine whether or not hospital patients had the identified Pneumonia disease. It provides update to probability of hypothesis (H) for some given instance of data (D) which can be expressed in equation 3.1 as follows:

$$P\ (H/D) = \frac{P\ (D/H) P\ (H)}{P\ (D)} \qquad (1)$$

Where P (D/H) is the probability of hypothesis and P (D) dataset features/parameters.

The character or feature variables are encoded using label encoder at preprocessing stage and feature scaling technique employed for the training and testing dataset of the independent variables in producing better classification report.

The D is given as:

$$D = (d_1, d_2, d_3, \ldots d_n) \qquad (2)$$

Where $d_1, d_2, d_3, \ldots, d_n$ represents the features mapped into the outlook.

Model Evaluation: It is necessary to use a standardized evaluation method in order to compare the various results. There are various methods of making correct and incorrect predictions for problems involving binary classification, such as the proposed system. Four different predictions are used to present the results. If the training sample provides the true value x and the model provides the prediction y. In this study, a match will be indicated with I if it is True, and a non-match will be indicated with 0 if it is False. As a result, the following criteria can be used to measure the model's performance in predicting the target variable:

The error rate can be measured with the general equation given by:-

$$\text{Error rate} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \qquad (3)$$

The sensitivity or also called the True Positive Rate (TPR) is given by:

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \qquad (4)$$

The False Positive Rate (FPR) or also called Fall-Out is gievn by:

$$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \quad (5)$$

$$\text{Accuracy} = \frac{\text{The total number of correct classification}}{\text{Overall number of classes}}$$

$$= \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Where TN represents true negative, FP is false positive, TP is true positive and FN is false negative cases.

Sensitivity is the ratio of the number of correctly classified positive classes of data computed using a function in Python as given in equation 3 as:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (7)$$

RF Algorithm 1: Random Forest (RF)

Step	**Processes involved**
1	Start
2	Assume cases in the training set to be C and randomly select cases with replacement.
3	If there are inputs of M features with a variable representing number m<M being specified such that at each node, "m" variables are randomly selected out of "M
4	Take the best split on the node and peg the value of "m" to be constant while we build-up the forest.
5	Grow each decision-tree to have the largest possible size without pruning
6	Predict with n_estimators using majority vote for classification and mean for regression.
7	Stop

Algorithm 2: Naive Bayes (NB)

Step	**Processes involved**
1	Start
2	**Input:**Training_Dataset (T)
3	F= (f_1, f_2, f_3,....f_n) // the predictor variables for testing items
4	**Output**: Class of testing items
5	**Compute** mean and standard deviation of predictor variables in each class
6	**Repeat** this step (a). Compute probabilities required for the Bayesian theorem for Exiting employees (b). Compute posterior probability of all those are not leaving the organization
7	**Compute** the likelihood of each class (first and second class)
8	**Get** the greatest likelihood
9	Return

RESULTS AND DISCUSSION

The necessary machine learning tools are used to present results and discuss findings of the proposed model. Improvements have been made to the concept and its application in order to produce better and more accurate results. In the following section, we presented and discussed the experimental findings of the proposed NB and RF-Based classifiers using confusion matrices, tree structures, tables, ROC curves and bar charts.

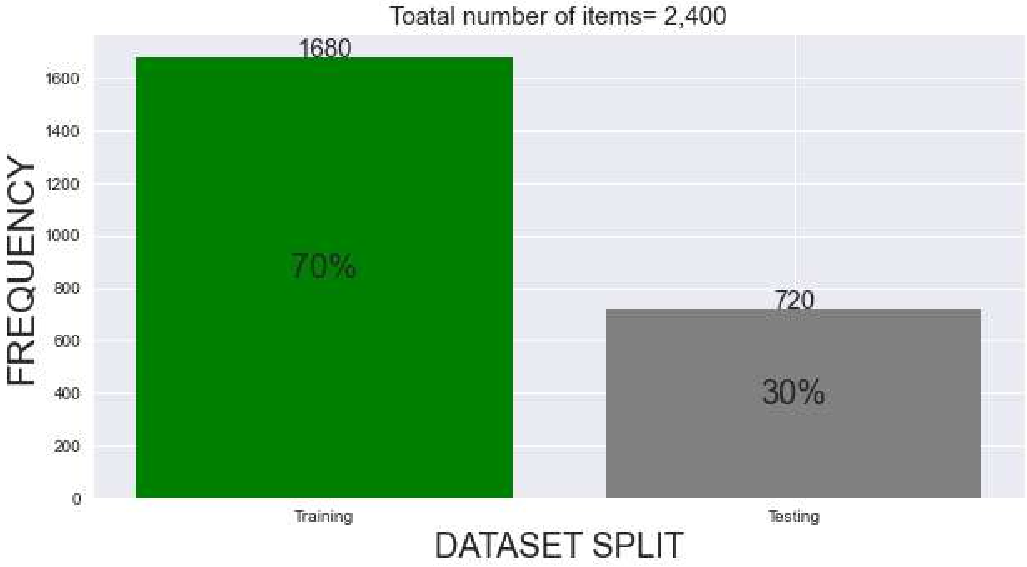


Figure 3 Training/testing data split

Figure 3 depcits the total dataset and tpercentage used for model training and testing. The total dataset used was 2400; of that, 70% (1680) were used for training, and the remaining 30% (720) were used for testing.

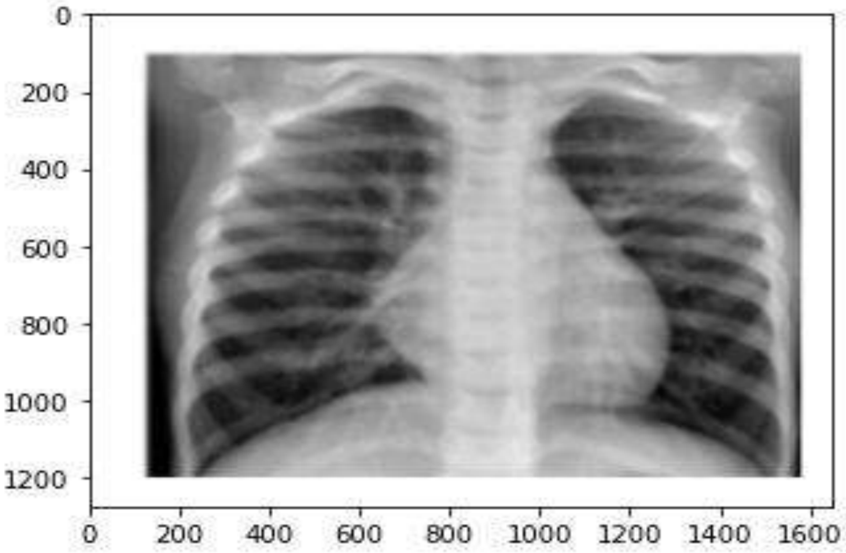


Figure 4: X-ray scan Image of healthy patients uploaded in python environment

Figure 4 shows an x-ray scan image of a patient without Pneumonia disease. A function was written in Python using the pillow and open computer vision (opencv) libraries to convert CX-ray scanned images into grayscale images.

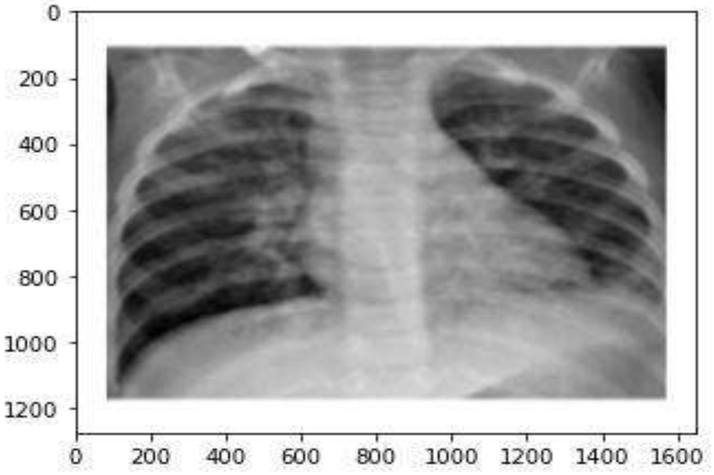


Figure 5: X-ray scan of patients who had pneumonia

Figure 5 depicts the scanned x-ray image of patients having features of Pneumonia diease infection. The x-ray test reveals infiltrates, which are white spots in the lungs that signify an infection. According to the x-ray scan results referenced above, the patient has pneumonia due to the excess fluids encompassing the lungs.

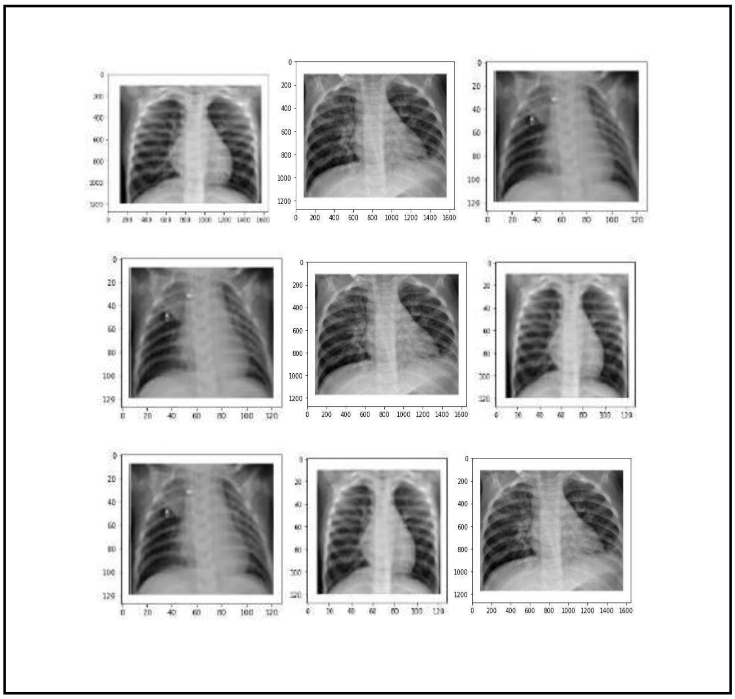


Figure 6: X-ray images

Patients with Pneumonia disease's x-ray images were gathered, preprocessed, and converted to grayscale images in the range of 1680 for training and 730 for testing.This was required in order to use Python's Open Computer Vision (opencv) library and create a more accurate clinical decision-making model.

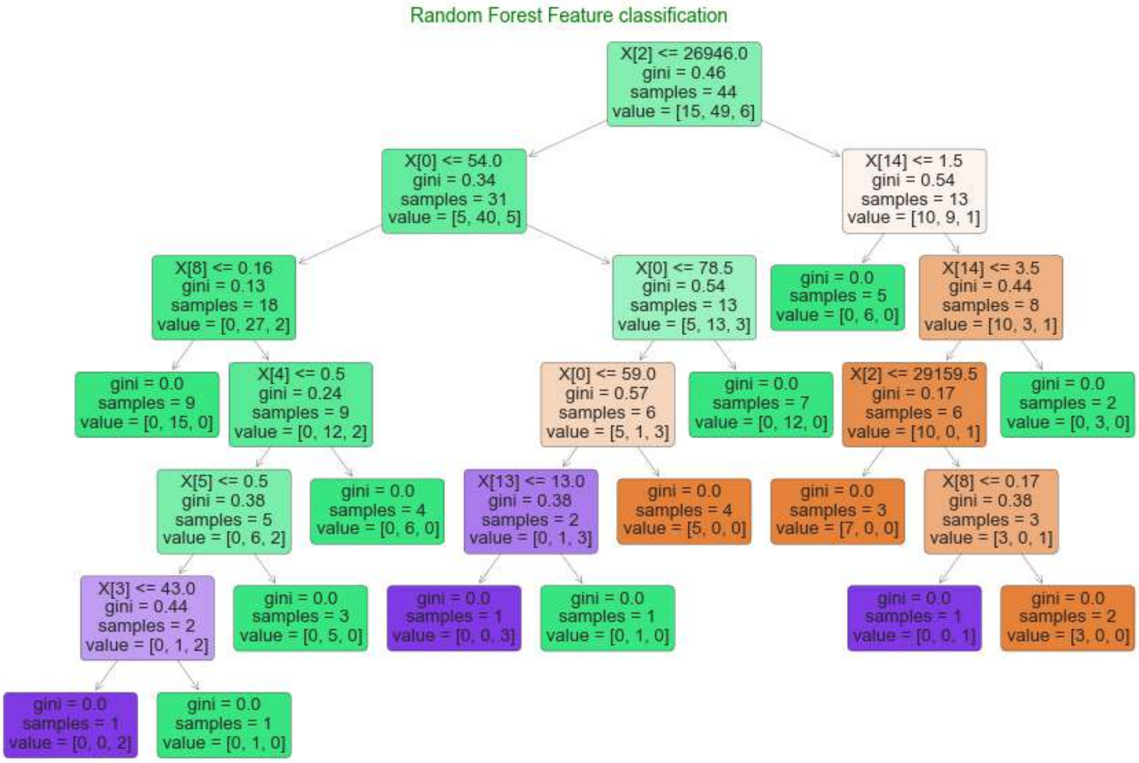

Random Forest Feature classification
X[2] <= 26946.0 gini = 0.46 samples = 44 value = [15, 49, 6]
X[0] <= 54.0 gini = 0.34 samples = 31 value = [5, 40, 5]
X[14] <= 1.5 gini = 0.54 samples = 13 value = [10, 9, 1]
X[8] <= 0.16 gini = 0.13 samples = 18 value = [0, 27, 2]
X[0] <= 78.5 gini = 0.54 samples = 13 value = [5, 13, 3]
gini = 0.0 samples = 5 value = [0, 6, 0]
X[14] <= 3.5 gini = 0.44 samples = 8 value = [10, 3, 1]
gini = 0.0 samples = 9 value = [0, 15, 0]
X[4] <= 0.5 gini = 0.24 samples = 9 value = [0, 12, 2]
X[0] <= 59.0 gini = 0.57 samples = 6 value = [5, 1, 3]
gini = 0.0 samples = 7 value = [0, 12, 0]
X[2] <= 29159.5 gini = 0.17 samples = 6 value = [10, 0, 1]
gini = 0.0 samples = 2 value = [0, 3, 0]
X[5] <= 0.5 gini = 0.38 samples = 5 value = [0, 6, 2]
gini = 0.0 samples = 4 value = [0, 6, 0]
X[13] <= 13.0 gini = 0.38 samples = 2 value = [0, 1, 3]
gini = 0.0 samples = 4 value = [5, 0, 0]
gini = 0.0 samples = 3 value = [7, 0, 0]
X[8] <= 0.17 gini = 0.38 samples = 3 value = [3, 0, 1]
X[3] <= 43.0 gini = 0.44 samples = 2 value = [0, 1, 2]
gini = 0.0 samples = 3 value = [0, 5, 0]
gini = 0.0 samples = 1 value = [0, 0, 3]
gini = 0.0 samples = 1 value = [0, 1, 0]
gini = 0.0 samples = 1 value = [0, 0, 1]
gini = 0.0 samples = 2 value = [3, 0, 0]
gini = 0.0 samples = 1 value = [0, 0, 2]
gini = 0.0 samples = 1 value = [0, 1, 0]

Figure 7 Random forest tree visualization

Figure 7 depict the forest tree genertaed from proposed system dataset with estimators properties containing an array of decision tree objects that made up the forest. To overcome the over-fitting issue, the forest tree generated from specific data sets is predicated on taking the majority vote (ranking) as the final result.

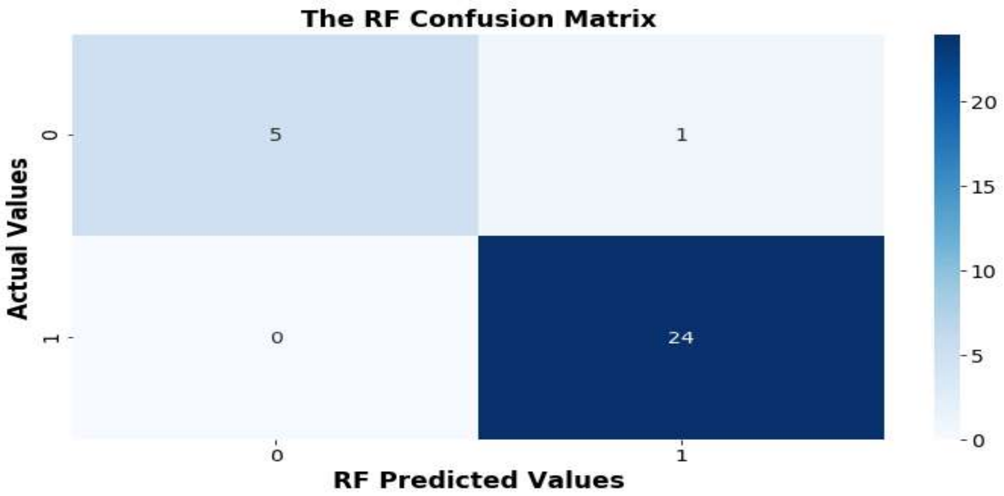


Figure 8: The confusion matrix of existing RF model

In figure 8: depicts the confusion matrix of the existing RF model with the correct predictions displayed at the secondary diagonal and wrongly predicted values recorded above and below the main diagonal called the off-diagonal elements. The total no. of correct predictions = TP+ TN =5+

24 =29 and wrongly predicted = 0 + 1 = 1 shown above in figure 4.2 where TP is true positive, FP false positive, FN false negative and TN true negative. The RF model encountered a type-I (FP=1) error.

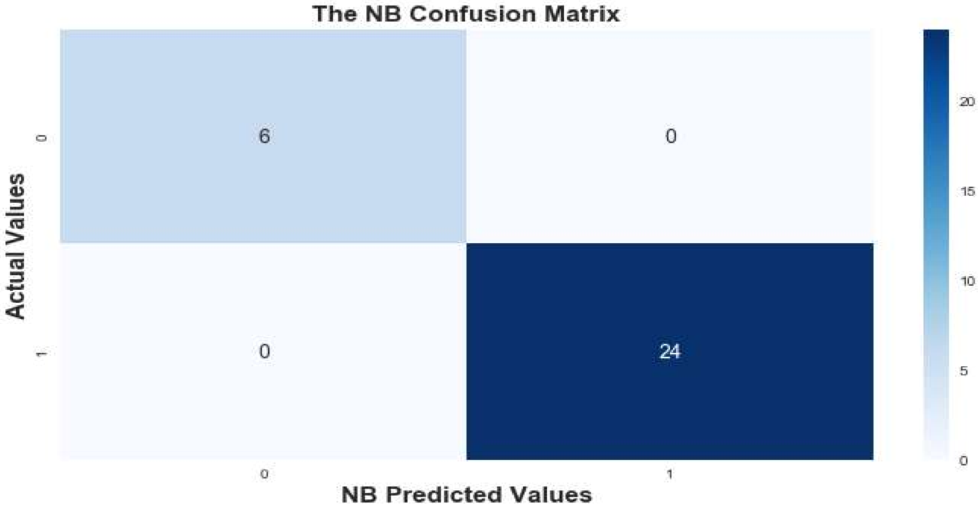


Figure 9: The confusion matrix of proposed NB model

Figure 9 depicts the confusion matrix of proposed NB classifier with leading diagonal elements or values showing the total number of correctly predicted values that are equal to the actual or true values above and below the main diagonal. The off-diagonal elements are the wrongly predicted values. The higher the diagonal values the better the recognition accuracy. From the confusion matrix: The total no. of correct predictions (TP+TN) =6 + 24 = 30 and wrong predictions (FP+FN) = 0 + 0 = 0.

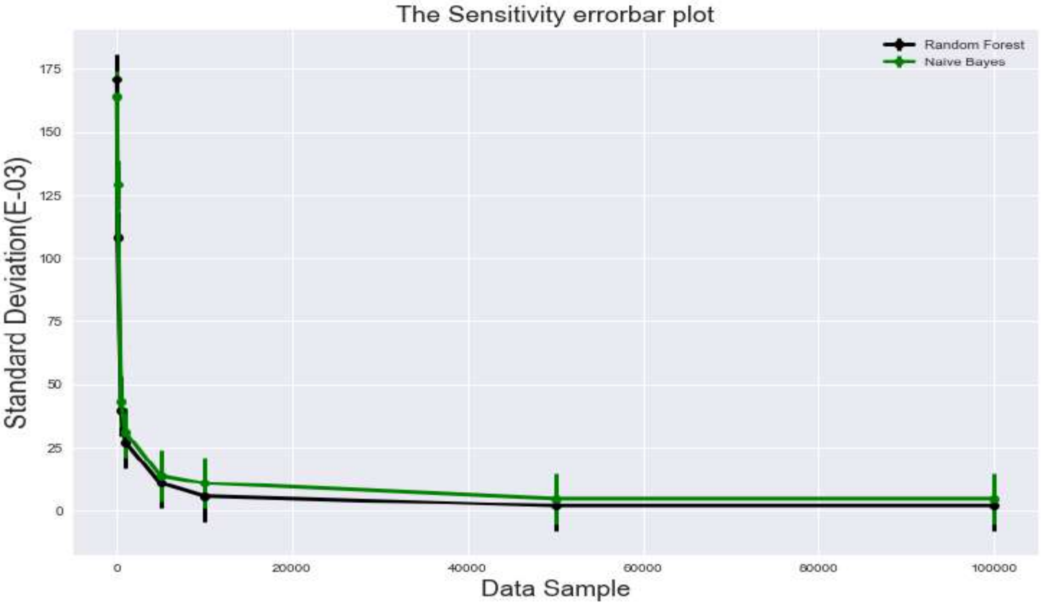


Figure 10: The sensitivity of plot of standard deviation

Figure 10 depicts the sensitivity analysis of NB and RF techniques used as a metrics to measure the variation of model's performances against different data sample in terms of standard deviation. The sensitivity value of NB classier against standard deviation recorded better improvement compared to RF over different data samples.

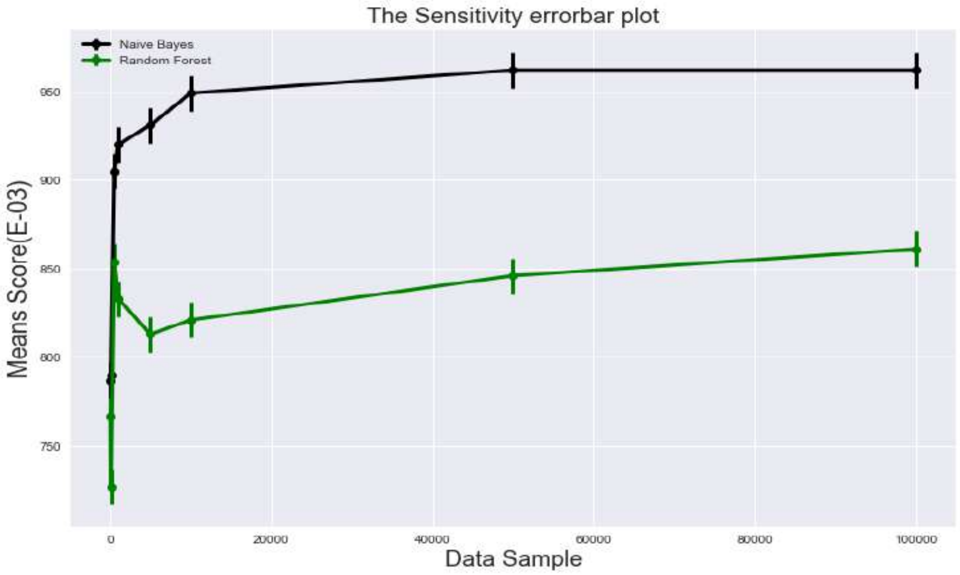


Figure 11: The sensitivity of plot of mean score

Figure 11 is the plot of RF and NB means score values against different data sample in measuring the performance of both models. There is a wide disparity and variation on the performance of NB and RF classifiers against different sample size. The sensitivity value of Naïve Bayes classification against the mean score recorded better improvement compared to Random Forestover different data samples.

Table 2: The multi-classification report of NB

Metrics	Pneumonia detected	Free cases
TP	5.000000	24.000000
TN	24.000000	5.000000
FP	0.000000	1.000000
FN	1.000000	0.000000
TPR	0.833333	1.000000
Recall	0.833333	1.000000
Sensitivity	0.833333	1.000000
TNR	1.000000	0.8333333
Specificity	1.000000	0.833333
FPR	0.000000	0.166667
PNR	0.166667	0.000000
PPV	1.000000	0.960000
Precision	1.000000	0.960000
F1-score	0.909091	0.979592

Table 2 depicts the multi-classification report of NB model for true positive (TP), true negative (TN), false negative (FP), recall, sensitivity, true negative rate (TNR), specificity, positive predictive value (PPV), and f1-score against detected and free cases of Pneumonia patients.Pneumonia affected patients produced 5.0000, 24, 0, 1, 0.833, 0.833, 0.833 and 1 cases of TP, TN, FP, TPR, Recall, sensitivity and TNR, whereas unaffected cases recorded 24, 5, 1, 0, 1, 1,1 and 1. Specificity, FPR, PNR, PPV, Precision, and F1-score values for pneumonia detected cases were 1, 0, 0.1, 1, and 0.909091, while those for undetected cases were 0.8333, 0.16666, 0, 0.96, and 0.979392.

Table 3: The multi-classification report of RF

Metrics	Pneumonia detected	Free cases
TP	6.0	24.0
TN	24.0	6.0
FP	0.0	0.0
FN	0.0	0.0
TPR	1.0	1.0
Recall	1.0	1.0
Sensitivity	1.0	1.0
TNR	1.0	1.0
Specificity	1.0	1.0
FPR	0.0	0.0
PNR	0.0	0.0
PPV	1.0	1.0
Precision	1.0	1.0
F1-score	1.0	1.0

Table 3 showcases the multi-classification report of the RF model with performance indicators of true positive (TP), true negative (TN), false negative (FP), recall, sensitivity, true negative rate (TNR), specificity, positive predictive value (PPV), and f1-score against detected and free cases of pneumonia patients. Patients with pneumonia had TP, TN, FP, TPR, recall, sensitivity, and TNR cases of 6, 24, 0, 0, 1.0, 1.0, 1.0, and 1.0, respectively, while unaffected cases had 24, 6, 0, 0, 0.0, 0, 1.0, 1.0, 1.0, and 1.0. Pneumonia cases that were detected had specificity, FPR, PNR, PPV, Precision, and F1-score values of 1.0, 0.0, 1.0, 1.0, and 1.0, whereas undetected cases had specificity, PPV, Precision, and F1-score values of 1.0 and 0.0, respectively.

Scientia Africana, Vol. 22 (No. 1), April, 2023. Pp 97-114 *https://dx.doi.org/10.4314/sa.v22i1.10*
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

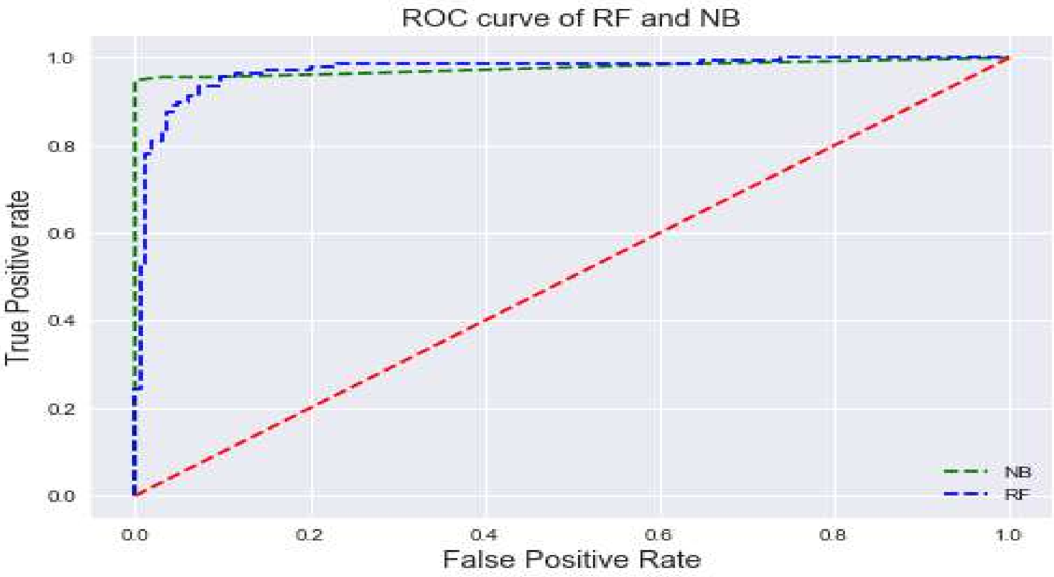


Figure 12: The ROC graph of existing RF and proposed NB

Figure 12 is the ROC curve of the RF and NB scores demonstrating how well the suggested models work. It is one of the most important performance indicators used to evaluate any classification algorithm's efficacy. With a relatively high AUC value, it shows how accurate the model is at differentiating between true positive and false negative data classes. The NB model's AUC curve was closer to the top left corner of the y-axis than the RF classifier's, which is a little farther away.

Table 4: Comparing NB and RF classifiers

	ALGORITHM	ACCURACY	RMSE
0	Naïve Bayesian (NB)	99.08	0.02
1	Random Forest (RF)	97.0	3.33

Table 4.3 depicts the accuracy and RMSE value of the proposed RF and NB prediction models as compared to the testing dataset. One of the most frequently used metrics for evaluating the precision of predicted results is the root mean square error (RMSE). The Euclidean distance concept's relative measurement error called the RMSE shows how much farther predictions deviate from calibrated true values. The most accurate metrics were produced by the NB model (100%) with no RMSE value, while the RF model produced the least accurate metrics (097.0%) with the highest RMSE (3.33) value.

CONCLUSION

The proposed model has proven to be extremely useful in real-world applications for quickly, consistently, and accurately diagnosing Pneumonia disease using chest X-rays. We draw the conclusion from the analysis above that the proposed system, with the addition of a hashing-based function, produced high detection accuracy as well as an improved classification report, f1-score, and precision. The Pneumonia disease target classes were successfully and accurately identified by the NB classifier without any misclassification errors.

It has been determined that a system is needed to assist in the diagnosis of the pneumonia disease based on the results of this proposed study. We therefore advise that this system be used by medical personnel, including radiologists and doctors of radiology, medical laboratory scientists, hospital patients, and AI and machine learning engineers.

Further Studies

This system was trained using exist dataset, something that may become out-dated in the future, but it can also be used for future testing in order to pass new test cases. Future work could take this paper to a relatively high level in order to enhance the system's changes and increase user confidence and other impacts of the model explanations. The Pneumonia disease detection system can also be enhanced with Explainable Artificial Intelligence (XAI) or Interactive Machine Learning (IML) techniques to help identify, locate, and correct Deep Learning model diagnostic errors with the aid of a user feedback mechanism.

REFERENCES

Abdullah, S. H., Abedi, W. M. S., and Hadi, R. M. (2022), Enhanced Feature Selection Algorithm for Pneumonia Detection, Periodicals of Engineering and Natural Science, 10 (6), 168-180.

Abdulkareem, N. M. and Abdulazeez, A. M. (2021), Machine Learning Classification Based on Radom Forest Algorithm: A Review, International Journal of Science and Business (IJSAB), 5 (2), 128-142.

Alsharif, R., Al-Issa, Y., Alqudah, A. M., Qasmieh, I. A., Mustafa, W. A. and Alquran, H. (2021) PneumoniaNet: Automated detection and classification of pediatric pneumonia using Chest X-ray images and CNN approach," Electronics (Basel), 10 (23), 2949.

Al-Dulaimi, D. S., Mahmoud, A. G., Hassan, N. M. Alkhayyat, A. and Majeed, S. A. (2022), Development of Pneumonia Disease Detection Model Based on Deep Learning Algorithm, Wireless Communications and Mobile Computing, 1-10, https://doi.org/10.1155/2022/2951168.

Alshehri, M. D., Alenazy, W. M., Hoang, T. V. and Alturki, R. (2021), Identification of Pneumonia Disease Applying an Intelligent Computational Framework Based on Deep Learning and Machine Learning Techniques, Edge Intelligence in Internet of Things using Machine Learning, 1-16.

Ash, B. and Turkington, C. (2007). The Encyclopedia of Infectious diseases (3 Ed) New York, 242.

Chattopadhyay, S., Kundu, R., Singh, P. K., Mirjalili, S. and Sarkar, R. (2022) Pneumonia Detection from Lung X-ray Images using local search aided sine cosine algorithm based deep feature selection method, International Journal of Intellectual System, 37 (7), 3777–3814.

El-Asnaoui, K. (2021), Design Ensemble Deep Learning Model for pneumonia disease classification, International Journal of Multimedia Inf Retr.10 (1), 55–68

Guleria, K. and Sharma, S. (2023), A deep learning based model for the detection of Pneumonia from Chest X-ray Images using VGG-16 and Neural Networks, Procedia Computer Science: International conference on Machine learning and data engineering, 28, 357-3666.

Gupta, P. (2021) Pneumonia Detection Using Convolution Neural Network, International Journal for Modern Trends in Science and Technology, 7 *(01), 77-80.*

Harshvardhan, G., Kumar, M. Rautaray, S. S. and Pandey, M. (2021) Pneumonia

Detection Using CNN Through Chest X-ray, Journal of Engineering Science and Technology, 16 (*1), 561-876.*

Ibrahim, Z., Bean, D., Searle, T., Qian, L., Wu, H., Shek, A., Kraljevic, Z., Galloway, J., Norton, S., Teo, J. T. and Dobson, R. J. (2017), A Knowledge Distillation Ensemble Framework for Predicting Short and Long-Term Hospitalization Outcomes From Electronic Health Records Data, Generic Colorized Journal, 2 (4), 1-14.

Kareem, A., Liu, H. and Sant, P. (2022), Review on Pneumonia Image Detection: A Machine Learning Approach, Human-centric Intelligence System, 2, 31-43.

Katoch, S., Sharma, S. and Singh, R. P. P. **(2021),** Pneumonia Disease Detection Using Deep Learning Methods from Chest X-Ray Images: Review, International Journal of Advanced Trends in Computer Science and Engineering, 10 (4), 2734-2740.

Liang, G. and Zheng, L. (2020) A Transfer Learning Method with deep residual network for pediatric pneumonia diagnosis, Comput. Methods Programs Biomed., 187 (104964), 104964.

Mabrouk, A., Díaz-Redondo, R. P., Dahou, A., Abd-Elaziz, M. and Kayed, M. (2022) Pneumonia detection on chest X-ray images using ensemble of deep convolutional neural networks," Applied Science. (Basel), 12 (13), 6448.

Mahapatra, D. (2014). Analyzing Training Information from Random Forests for Improved Image Segmentation. IEEE Transactions on Image Processing, 23 (4), 504-1512.

Masud, M., Bairagi, A. K., Nahid, A. A., Sikder, N., Rubaiee, S., Ahmed, A. and Anand, D. (2021), A Pneumonia Diagnosis Scheme Based on Hybrid Features Extraction from Radiographs Using an Ensemble Learning Algorithm, Journal of Healthcare Engineering, *1-11.*

Mcluckie, A. (2009). Respiratory Disease and its management DOI: 10.1007/978-1-095-1.

Panjasuchat, M. and impiyakorn, Y. (2020). *Applying Reinforcement Learning for Customer Churn Prediction.* Journal of Physics: Conference Series, 1619 (1), 12015

Prayogo, K. A., Suryadibrata, A., and Young, C. J. (2020) Classification of Pneumonia From -ray Images Using Siamese Convolutional Network, TELKOMNIKA Telecommunication, Computing, Electronics and Control, 18 (*3), 1302-1309.*

Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, AartiBagul, Curtis P. Langlotz, Katie S. Shpanskaya, Matthew P. Lungren and Andrew Y. Ng. (2017) CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, 1-7.

Raj, T. F. M. and Prasanna, S. (2012) Implementation of ML Using Naive Bayes Algorithm for Identifying Disease-Treatment Relation in Bio-Science Text, Research Journal of Applied Sciences, Engineering and Technology, 5 (*2), 421-426.*

Ramakrishnan, R., Bhattacharya, S. and hanya, P. (2018). Predict Employee Attrition by Using Predictive Analytics, *Benchmarking: An International Journal, 26 (1), 2-18.*

Stephen, O., Sain, M., Maduh, U. J. and Jeong, D. U. (2019) An Efficient Deep Learning approach to pneumonia classification in healthcare, Journal of Health. Engineering., 4180949.

Varshni, D., Thakral, K., Agarwal, L., Nijhawan, R., and Mittai, A. (2019), Pneumonia Detection Using CNN based Feature Extraction, 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 1-20.

Wang, J., Liu, W., Kumar, S. and Chang, S. (2015) Learning to Has for indexing Big Data – A survey, Proceedings of the IEEE, 1-11.