INTERNATIONAL JOURNAL OF
Telecommunications
AIR DEFENSE COLLEGE

Improving Life-threatening Lung Diseases Classification using Hybrid SMOTE-ENN with assorted Machine Learning Classifiers

Citation: Albanhawy , M.; Khalil , A.; Moustafa , H. Title.

*Inter. Jour. of Telecommunications, **IJT'2024**, Vol. 04, Issue 02, pp. 1-18, 2024.*

Editor-in-Chief: Youssef Fayed.

Received: 07/07/2024.

Accepted: date 17/08/2024.
Published: date 17/08/2024.

Publisher's Note: The International Journal of Telecommunications, IJT, stays neutral re-gard-ing jurisdictional claims in published maps and institutional affiliations.

Mostafa Mahmoud Albanhawy [1, *], Abeer Khalil [2], and Hossam Moustafa [3]

* Correspondence: engineer.mostafa.albanhawy@gmail.com; M.Sc. student at the Department of Electronics and Communications Engineering at the Faculty of Engineering, Mansoura University
[2] Associate Professor at the Department of Electronics and Communications Engineering, at the Faculty of Engineering, Mansoura University; abeer.twakol@mans.edu.eg
[3] Professor at the Department of Electronics and Communications Engineering at the Faculty of Engineering, Mansoura University; hossam_moustafa@mans.edu.eg

Abstract Chest radiography is one of the most common diagnostic tools for diagnosing and managing bronchopneumonia and other lung diseases. In this paper, a classification strategy was proposed for identifying infection in Chest X-ray images. We collected 7545 x-ray chest images from an openly accessible X-ray database and separated them into three classes: healthy individuals, persons suffering from pneumonia, and additional COVID-19 patients. The contrast limited adaptive histogram equalization (CLAHE) method was used to improve the quality of the X-ray images. The oriented gradient histogram (HOG) is used. The classification of datasets in medicine sometimes is hindered by the problem of having unequal datasets. In the solving of this problem, which occurs during imbalanced data classification in medical diagnosis, we introduce a hybrid sampling technique called SMOTE-ENN that is a combination of the Synthetic minority oversampling technique (SMOTE) and Edited Nearest Neighbors (ENN). The support vector machine (SVM), k-Nearest Neighbors (k-NN), and Random Forest Classifier (RFC) used to classify the images, with classification rates of 99.47%, 98.70%, and 98.47%, respectively, on a test dataset of 1504 images. These findings may help to detect COVID-19 and pneumonia diseases more effectively.

Keywords: Pneumonia, Covid-19, CLAHE, SVM, k-NN, CXR, Oriented Gradients Histogram, SMOTE, ENN

1. Introduction

The significance of disease classification and prediction has become increasingly clear in recent years. To accurately identify the exact cause and symptoms of a disease, it is essential to have a comprehensive understanding of the important properties and features of a dataset [1]. Artificial Intelligence (AI) has shown its ability to classify diseases and assist decision-making processes [2]. Machine Learning (ML), a subset of AI, has significantly accelerated research in the medical field. By leveraging data, ML has enabled advancements in various areas, such as natural language processing, automatic speech recognition, and computer vision, resulting in the development of robust systems like driverless cars and automated translation [3]. Despite these remarkable advancements, the application of ML in medical practice has encountered certain challenges and risks. The medical community has raised concerns regarding the goal of making accurate predictions using collected data and the effective management of such predictions within the medical system [4]. Classification algorithms are extensively used in hospitals to precisely categorize diseases, particularly in the screening and prevention of various life-threatening thoracic diseases [5]. In the medical diagnostics field, the task of multi-label classification poses significant challenges. The lungs can be examined using radiography (X-rays), computed tomography (CT), and magnetic resonance imaging (MRI), which are the three main radiological methods used [6]. In recent years, the prevalence of various factors contributing to lung diseases has led to increased mortality rates [7]. Individuals affected by COVID-19 and pneumonia exhibit mild to moderate symptoms, such as fever, coughing, and dyspnea. However, some individuals develop severe pulmonary complications in their lungs, ultimately resulting in fatality as well [8]. X-ray radiographs serve as a cost-effective and non-intrusive technique for investigating various bodily organs. It is estimated that approximately 3.6 billion X-ray images are captured annually on a global scale. When radiologists evaluate CXR images for longer periods, the risk of making the wrong diagnosis increases. This illustrates the significance of efficiently and in a timely manner in preventing wrong diagnoses in patients [9]. The United States alone has conducted more than 150 million chest X-ray radiographs (CXR). X-rays can track disease progression throughout the treatment course [10]. No matter the essential function that CXRs play inside the analysis of chest sicknesses, visible interpretation with the aid of radiologists stays challenging and can cause blunders [11]. In addition, even radiologists with experience are more likely to misdiagnose when subtle forms of pathology exist in the hamstrings or some inner organs than they know from a chest radiograph reading entirely different statements about what they have seen there on screen images produced under scope field of vision [12].Moreover, the World Health Organization (WHO) warns that many chest diseases could kill people if not attended to properly, with life-losses significant and potentially affecting the lives of tens or hundreds millions, in case millions die [13]. Some chest diseases, such as tuberculosis, claiming approximately 1.4 million lives annually, pneumonia causing the deaths of 9 million children under 5 years old as the leading killer disease worldwide, and COVID-19 responsible for over 6 million deaths globally as of November 2022, have high mortality rates [14]. Numerous studies have been conducted to enhance classification accuracy by developing sophisticated classification approaches and methods. The deficit of radiologists in multiple countries globally remains a significant issue, particularly given the large number of individuals requiring radio-logical examinations, which exceeds the rate at which new radiologists can be trained. The extended wait times for diagnostic procedures at medical institutions, the excessive incidence of medical error of chest X-ray images, and the prevalence of severe sicknesses underscore the importance of developing good computer-assisted diagnostic systems for early detection of chest diseases. [15]. With this objective in mind, CAD was created with the aim of efficiently assisting radiologists in achieving high-quality diagnostic results, thereby improving patient care. CAD systems are designed to provide a valuable second opinion that complements the expertise of radiologists rather than aiming to replace or challenge their role. In healthcare, diagnosis is a critical step, ensuring that patients receive appropriate treatment for their specific illness as soon as possible. Machine learning techniques analyze medical data to forecast the health status of individuals. Owing to the massive volume and complexity of the data, manual analyses by physicians are out of the question in large data sets regarding how one can accurately predict diagnosis. By harnessing the power of machine learning, patient data can be effectively scrutinized, enabling informed predictions concerning the likelihood of certain diseases. [16].

Chest radiography is a common diagnostic modality that does not involve any invasion or pain. It is primarily used for photographing the lungs, heart, windpipe, and bone structure of the chest. A posterior-anterior view, we would see that air shows-up black while bones come out white, and between these at intermediate levels with regard to the amount of tissue remaining, which is something in place. Because healthy lungs are full of air, the lung area will appear darker in contrast to surrounding bones and tissue. Also, the standard chest radiograph of a healthy person shows different prevalent attributes, such as well-defined cost phrenic angles, a marked hemi diaphragm line in its entirety, and distinct edges around organs including heart etc. [17].
Specific radiographic features in chest X-rays may allow for the identification of pneumonia in patients. Among these is the evidence presented by dull cost phrenic angles regarding pleural effusion, areas in the lung fields with white or misty shadows, indefiniteness around the heart borders, as well as corresponding signs within other sections of the body. This is proved by reference [18].

Healthcare is a pivotal sector in human existence and economic structure. Extensive evidence has substantiated the positive correlation between a thriving healthcare sector and a prosperous economy [19]. The healthcare sector has witnessed remarkable advancements in recent years due to the significant role played by artificial intelligence. This can be attributed to the availability of extensive clinical data that serve as the basis for training these systems. As a result, these data can be used to detect, predict, and discover optimal treatments for various diseases. Studies have demonstrated that cough, fever, chest pain, pneumonia, and dyspnea are commonly observed during the early stages of the illness. For example, diseases such as asthma considerably affect the airways of the lungs, leading to inflammation and thus, breathing problems. Furthermore, pneumonia, tuberculosis, and lung cancer target air sacs called alveoli located in our lungs, correspondingly [20]. Alveoli become filled with fluid and pus in cases of pneumonia, an infectious disease. As a consequence, patients with this condition may experience shortness of breath and/or discomfort while breathing [21]. In 2019, Wuhan, China developed as the center of the Coronavirus disease 2019 epidemic, which is a highly infectious airborne disease that has since spread very rapidly across all continents and infected an astonishing number over 600 million individuals, with a death toll exceeding 6.5 million people tragically by September 2022 [22].

In medical datasets, there is a well-known issue in machine learning on data imbalance, which means that one class, may contain much more or much fewer instances than other classes [23]. An imbalanced data distribution can create obstacles for deep learning algorithms because they may exhibit a bias toward the majority class. Consequently, this could result in less than optimum performance on the minority class like reduced accuracy or poor sensitivity [24]. Multiple strategies are available to address data imbalances in medical datasets. One strategy involves oversampling the minority class, which can be achieved by generating synthetic examples or duplicate existing examples. Another strategy involves sampling the majority class, where certain examples from the majority class are removed [25].

In the ongoing paper machine learning methodologies designed to tackle pneumonia and COVID-19; two respiratory conditions that affect lung air sacs are outlined. The objective is to detect diseases at an early stage, thereby improving survival or successful treatment and supporting radiologists in their decision-making processes.

The paper's key contributions are given below:

- The objective is to design an accurate machine learning model that has undergone training using the same number of instances for each class—and this will enable radiologists to detect lung diseases early enough.
- We used a public combined curated dataset for Posterior-Anterior Chest X-ray images obtained by collating 15 publicly available datasets. The models were developed to perform a multi-classification task, which was to differentiate chest X-ray images of individuals with normal, pneumonia, and COVID-19 conditions.
- Employed a successful pre-processing approach to enhance chest X-ray images and make them more suitable for image analysis before feeding them to the feature extraction stage.
- We analyzed the chest images by applying an oriented gradient histogram.
- To address the class imbalance issue, model performance was improved by balancing the dataset using the synthetic minority oversampling technique and edited nearest neighbor.
- We trained three popular supervised machine learning algorithms: SVM, k-NN, and RFC.
- To find the best model, SVM achieved an overall accuracy of 99.47%, as well as precision, recall, specificity, f1 score, MCC, CSI, and CE of 99.19%, 99.55%, 99.73%, 99.37%, 99.11%, 98.75%, and 0.36, respectively. The results obtained by the proposed model are superior to those of other studies in the literature.

The remaining parts of the manuscript are organized step by step: In the second section, we will review previous studies. We will introduce the dataset and its characteristics while describing the methodology in the third section, and finally, we will report our findings in the fourth section. Section five presents the conclusion of this paper.

2. Literature Review

Several studies have examined X-ray images and come up with ways of diagnosing COVID-19 and pneu-monia.. This technology can be easily integrated into healthcare and has demonstrated remarkable potential for detecting COVID-19 and pneumonia. The diagram in discern one illustrates the structure of the chest area as showed in normal PA chest X-ray scans.

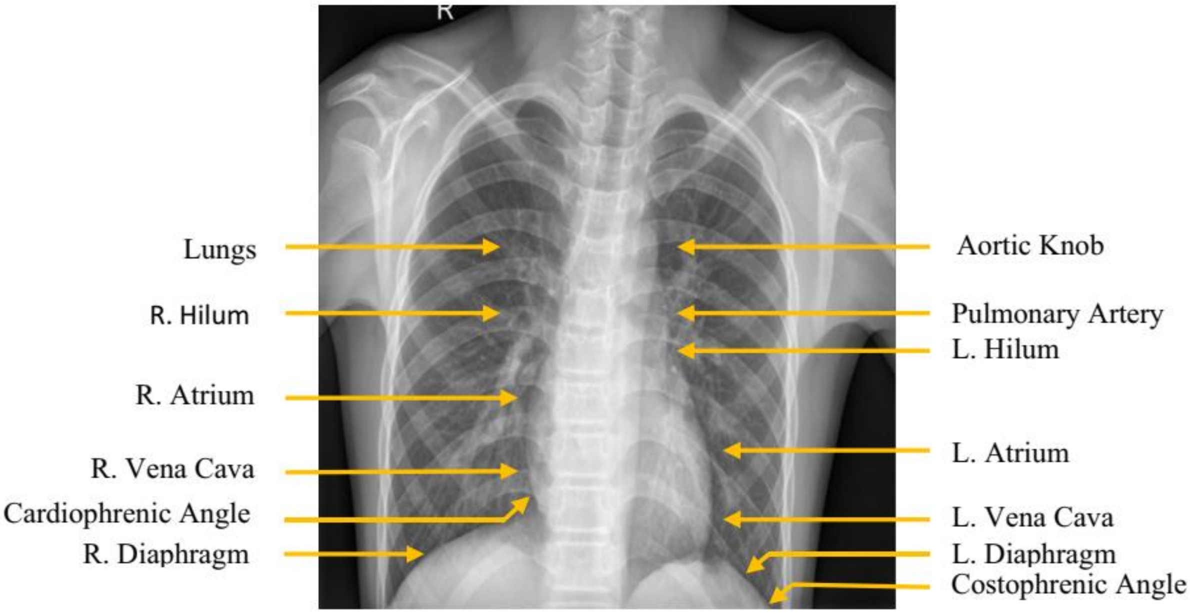
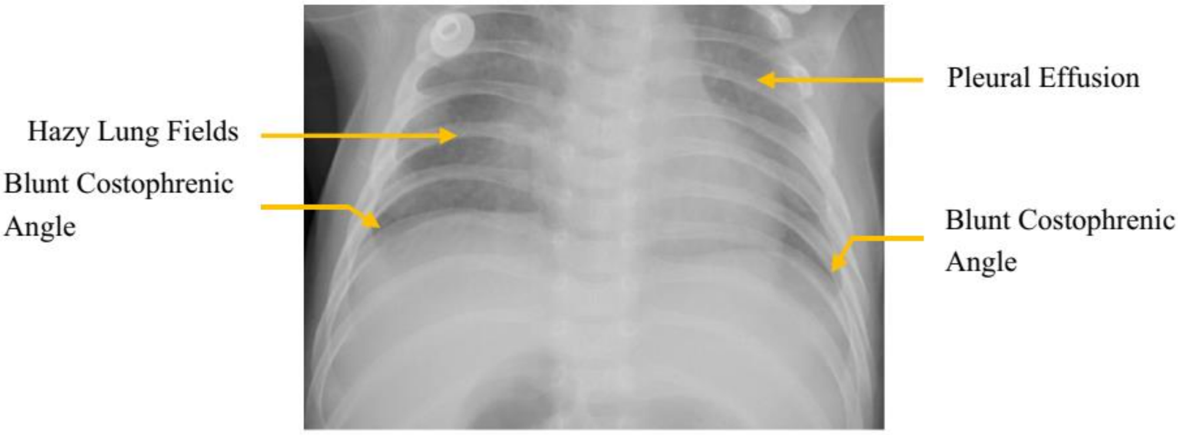


Figure 1. The chest X-ray reveals the presence of normal pulmonary artery anatomy, suggesting the absence of pathological findings.

Figure 2 shows multiple visually represented radiological signs of Pneumonia.



.

Figure 2. Poster anterior chest X-ray showing signs of pneumonia.

Several worthy contributions to this connection are listed below.
In [26], Yadav and Jadhav utilized a support vector machine (SVM) as a classification approach to classify pneumonia data into the following three classes: viral, normal, and bacterial pneumonia. They used deep learning approaches, such as VGG-16 and InceptionV3 models. The accuracy of their results was 96.6%. In [27], Stephen et al. developed a model to classify the presence of pneumonia from a collection of chest X-ray image samples, and the results obtained were training accuracy = 0.9531 and validation accuracy = 0.9373.

Zagrouba et al. [28] proposed a supervised machine learning model for simulating the COVID-19 outbreak by using a support vector machine (SVM) for forecasting. During training and testing, they attained 98.88% and 96.79% accuracy levels, respectively.

Rahman et al. [29] proposed a predictive model for the COVID-19 outbreak based on supervised machine learning. The results of their study on the SVM-based multi-fold cross-validation method produced good results, with an accuracy rate in the validation set of 98.4%.

In order to evaluate deep learning efficiency, Nagi et al. [30] used a large dataset, and they found that the kind of information that worked for them was the Xception model, which demonstrated a 94.21% accuracy rate compared to any other model. Alternatively, researchers whose findings have been published formally maintained that their particular customized model did not perform well compared to others that only recorded 92.38%.

Khuzani et al. [31] proposed a two-hidden-layer machine learning classifier (using these features formed) capable of identifying COVID-19 patients through chest X-ray scans. The dataset contains 420 chest X-ray images with (512×512) do these image are divided as 140 image for COVID-19, 140 image for pneumonia and 140 image for normal people. A: The model had a very high accuracy, precision, of 96%, 100%, and 0.98, respectively, for the COVID-19 class due to a relatively small training dataset.

To make these well-known deep convolutional neural networks applicable to low computing resources, a truncation method was proposed for reducing the parameters of the models [32] and implemented by Montalbo. The results demonstrated that the Inception-ResNet-V2 model when trimming and parameter crunching down to 441 K had an accuracy of 97.41% for three-class classification (Nor mal, COV ID-19, Pneu monia).

A model for normal, COVID-19, and pneumonia chest radiographs was proposed by Haque et al. [33]. They applied a fusion of ResNet-101 and ResNet-151 with a better dynamic weight ratio to enhance the model. During testing, the model demonstrated an accuracy of 96.1 %.

Wang et al. [34] introduced the COVID-Net with CNN model, which is a new model developed by them. to act as a baseline for COVID-19 diagnostics from chest radiographs obtained from the COVIDx dataset. With few lightweight residual patterns, the COVID-Net architecture was designed, and images were classified into three classes: COVID-19, pneumonia, and normal. The model was tested using an independent set of 1000 images, and it achieved 93.5% sensitivity at a specificity of 95.7%. This allows us directly compare how well COVID-Net is with levels for different styles of positive and negative COVID-19-identified cases embedded by these metrics.

According to Ragab and Mahmoud [35], a study was conducted using a capsule neural network model that included 6310 chest images. It was found that the CapsNet model achieved classification accuracy above 95 % in three classes; "Normal, COVID-19, and pneumonia".

3. Material and Methods

In this section, we describe the datasets of images used in this study. Then, we proceed to data preparation and processing. Then, we describe the feature extraction process, which is based on "a feature descriptor used for computer vision and analysis of medical images". Following this, a hybrid sampling algorithm that combines synthetic minority oversampling and edited nearest neighbor sampling is proposed as a solution to the problem of sample imbalance in medical datasets. Next, we present the classification techniques. Finally, we define the metrics used to evaluate the results and compare them to existing approaches. As proven in Figure 3, the proposed model is described in the following subsections.

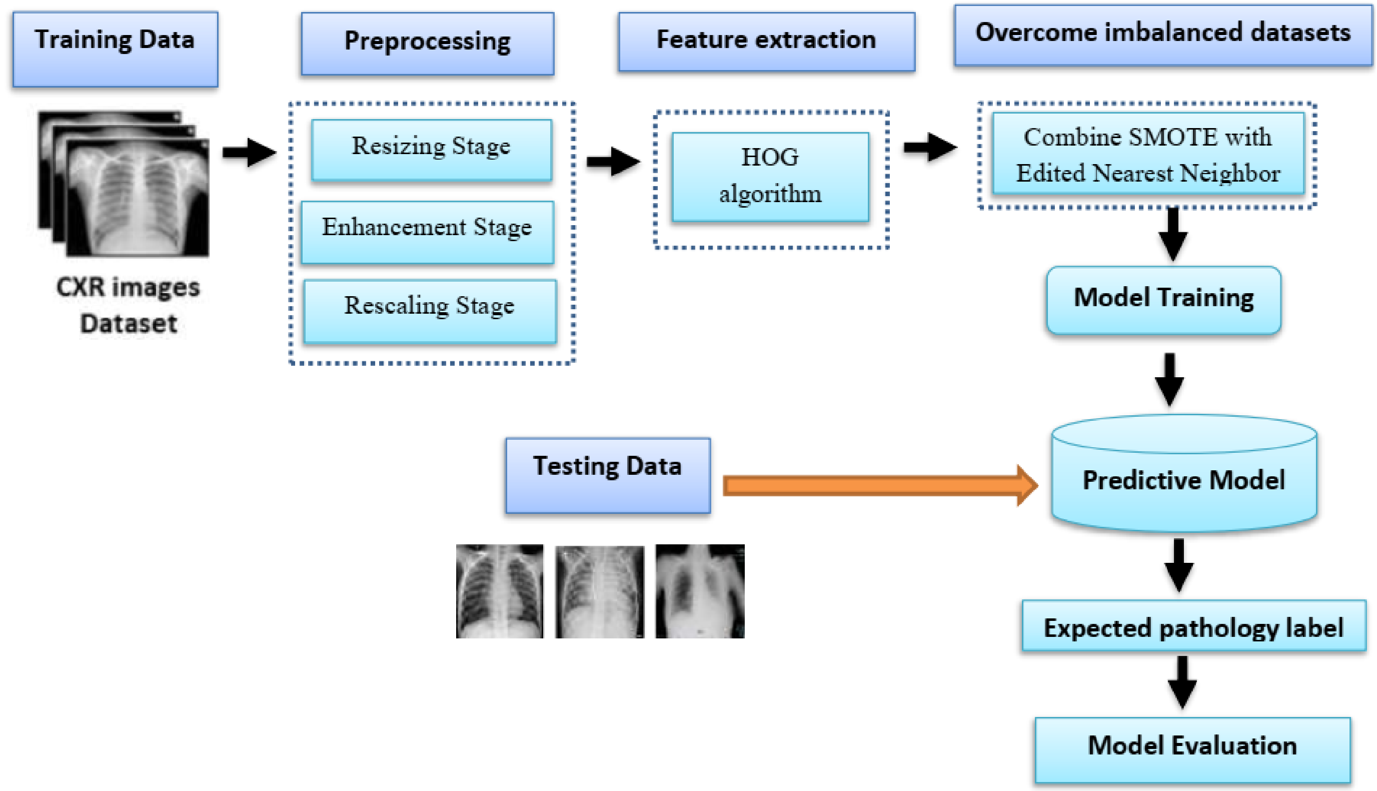


Figure 3. Structure of the proposed technique for lung disease classification

3.1. Chest X-ray Data

Data are an important input to machine learning and deep learning problems. A relevant collection of data helps analyze machine learning problems better. Data were collected from public sources and through agreements with hospitals and physicians with the consent of the patients. We collected data for this re-search from the Kaggle website [36], which includes chest X-ray images taken in the posterior-anterior (PA) view, which is available for researcher purposes. All images from the datasets were stored in the joint photographic experts group (JPG/JPEG) format. In this study, the dataset contains a total of 7545 chest X-ray images categorized as follows: 1275 cases of COVID-19, 3270 under normal conditions, and another 3000 X-rays showing other pneumonia cases. Figure 4 shows the distribution of different classes of X-ray images.

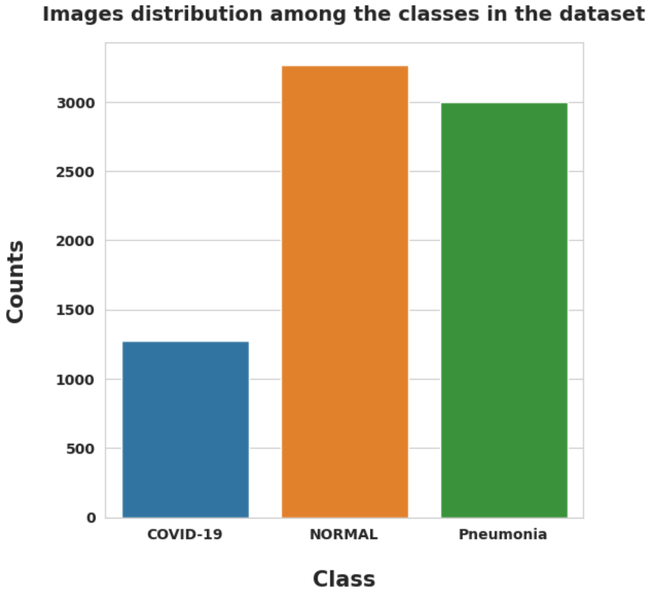
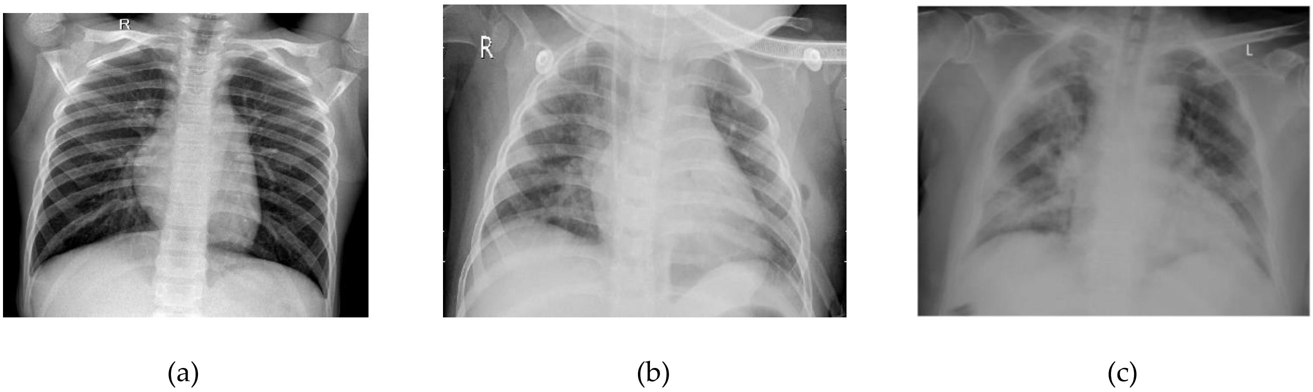


Figure 4. Distribution for different classes of Posterior-Anterior Chest Radiography Images.

3.2. Data Preparation and Preprocessing

Throughout this study, the dataset was randomly divided, and 80% of the data belonged to the training set, while the remaining 20% was used in the test set. Our experimental dataset contains 2616 healthy people, 2400 patients with pneumonia, and 1025 patients with COVID-19 are included in the training directory. The test directory contains images from 654 healthy people, 600 patients with pneumonia, and 250 patients with COVID-19 on chest X-rays. Figure 5 presents the X-ray dataset images showing various anteroposterior views.



(a) (b) (c)

Figure 5. Random PA view from each class. Normal a, Pneumonia b, COVID-19 c

Preparing raw data before using machine learning algorithms is known as preprocessing. The feature ex-traction quality and image analysis outcomes would be improved by performing image preprocessing. Poor category results emerge when we use original raw images in training machine learning algorithms. Among the primary tasks in image pre-processing are changing the size of the image, improving it, and scaling the information.

3.2.1. Resizing Stage

To overcome the computational limitations, the input images were downsized at this stage. The images were processed into grayscale as the initial step in the preprocessing process. Following this, the images are resized to dimensions of (144 × 144). The images were resized to 144 × 144 dimensions to match the smallest size available in the datasets used in this study.

3.2.2. Enhancement Stage

Image enhancement is a process used to remove unwanted distortion due to deterioration in contrast, unwanted noise, improper intensity saturation, blurring effect etc., and to determine hidden information contained in images. However, X-rays may also display images that lack contrast and are dark. Owing to these restrictions, the development of digital image enhancement technology is more popular among investigators and medical staff because it is necessary to extract useful information from these pictures and make them more readable.

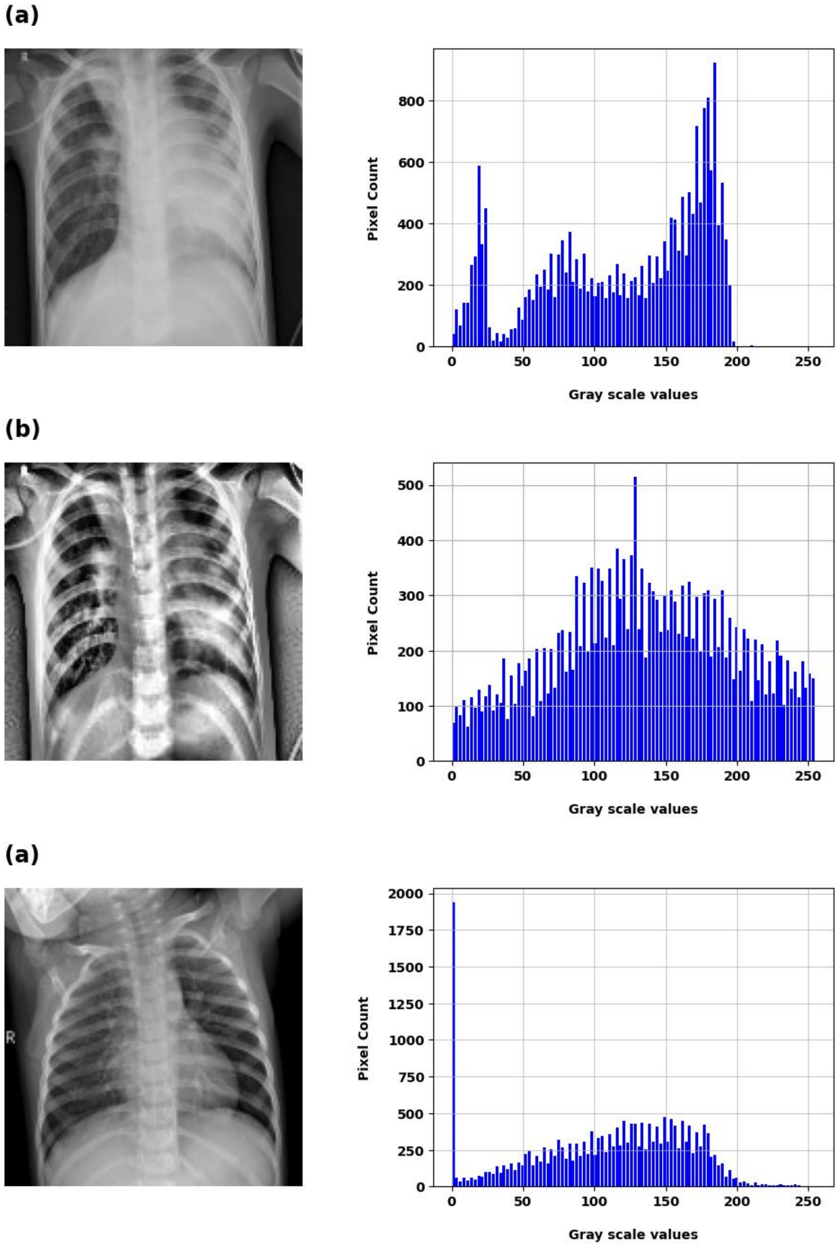
In image representation, pixels may have different values, such as a range of 0 to 255 for uint8 images with an interval [0, 1] for floating-point images. Nevertheless, it should be noted that quite a few images seem to carry a range of narrower values, often because of poor contrast. Alternatively, a large part of the pixel values in an image may be confined to a particular sub-range within all available code values. To enhance the visibility of images, this study utilized contrast limited adaptive histogram equalization. Improving contrast in images is the issue on which the CLAHE technique has been found to be successful, while enhancing the most common intensity value close to the peak intensity or extending the intensity distribution over the entire range can be achieved by using histogram equalization as a simple technique for image contrast enhancement [37]. CLAHE is almost similar to AHE except that histogram amplification is restricted by clipping out some values before computing the cumulative distribution function.

The over-amplified part of the histogram is redistributed over the histogram. Equation (1) shows the CLAHE computation as follows [38]:

$$p = \left(p_{(max)} - p_{(min)}\right) * p_{(f)} + p_{(min)} \tag{1}$$

Where, p represents the pixel value after applying CLAHE and $p_{(max)}$, $p_{(min)}$ represents maximum and minimum pixel value of an image and $p_{(f)}$ represents cumulative probability distribution function.

The results of this stage on Figure 6 showing X-ray images through which; There are figures showing input images, enhanced images, and histogram plots. Histogram plots appear dense in the original image pixel probability but relatively scattered in the enhanced image pixel probability, indicating improvement in image contrast.



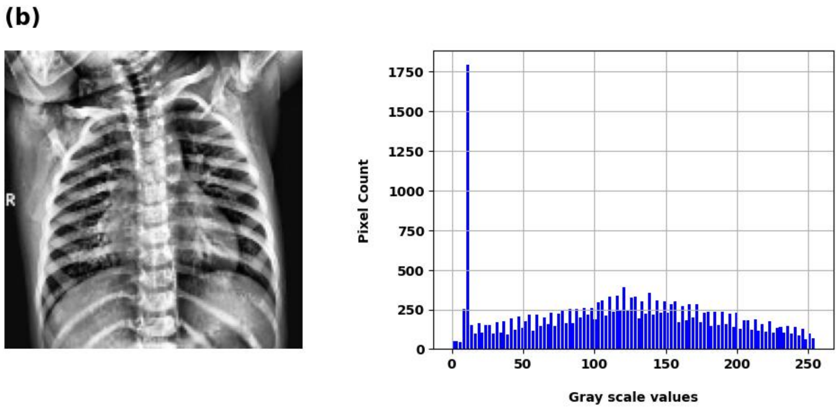


Figure 6. Shows how improving the X-ray image by using histogram analysis ends up as displayed by some events. (Original: Image (a); Enhanced: Image (b)).

In the source images, the contrast is stretched, but in the enhanced images, it is more uniform, with the pixel intensity evenly distributed throughout the pixels. The histograms clearly show that the pixel probability distribution in the images is concentrated.

3.2.3. Data Rescaling Stage

Rescaling is required to enhance the speed and accuracy and produce good results. The standard scaler is a method to preprocess data for machine learning. The standard scaler helps obtain standardized distribution. This technique standardizes features by subtracting the mean value from each feature and subsequently dividing the obtained result by the standard deviation of the feature [39]. In this study, we utilized the Standard Scaler method to preprocess the entire dataset. The standard scaling was calculated as [40]:

$$X_{new} = \frac{X_i - X_{mean}}{X_{std}} \tag{2}$$

Where, X_{new} is scaled data, X_i is to be scaled data, X_{mean} the mean of samples, X_{std} is the standard deviation of samples.

3.3. Feature Extraction

Histogram of Oriented Gradients is used to extract features from image. Dalal and Triggs used an oriented gradient histogram for pedestrian detection. It uses a gradient calculation with an image dimension of 128 height pixels and 64 width pixels [41].

Gradients (x and y derivatives) of an image are useful because the magnitude of gradients is large around edges and corners. The following definition applies to the gradient of the image f(x, y):

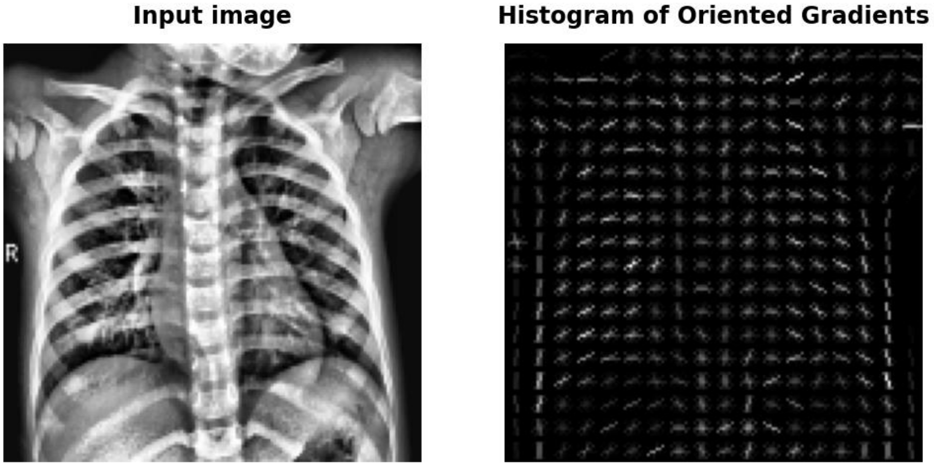
$$\nabla f = \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{3}$$

$\frac{\partial f}{\partial x}$ The image's derivative in terms of x, $\frac{\partial f}{\partial y}$ the image's derivative in terms of y. The kernels are used to convolve the images along the x and y axes for calculating the derivatives $\begin{bmatrix} -1 & 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$. Here are formulas that can tell us how to find a gradient's direction as well as magnitude :

$$\text{Magnitude of Gradient } (g) = \sqrt{g_x^2 + g_y^2} \tag{4}$$

$$\text{Direction of Gradient } (\Theta) = \arctan(\frac{g_y}{g_x}) \quad (5)$$

Hence, the total number of features for the image was 2916. Figure 7 shows a chest X-ray image from the dataset. In addition, there are histograms that show the computed oriented gradients of this image.



(a) **(b)**

Figure 7. Illustrate the features that have been extracted from an x-ray image. (Image enhancement: Image (a), Apply the HOG algorithm: Image (b)).

3.4. Handling class imbalance in chest X-ray images classification

ML classification algorithms typically assume equal samples for every class [42]. Class imbalance is when the distribution of classes in the dataset is unequal. If we were to train a classification algorithm over imbalanced data, we could presumably obtain predictive models that are not effective and have trouble accurately identifying the minority class. In order to solve this problem, this study used a hybrid sampling algorithm that combines the synthetic minority oversampling technique (SMOTE) and edited nearest neighbor (ENN). SMOTE-ENN Method Developed by Batista [43]. This can improve the generalizability of the model. Table 1 depicts the distribution of classes before and after balancing.

Table 1. Number of images in each class before and after handling class imbalance using SMOTE-ENN.

Class	Number of X-ray images in normal distribution	Number of X-ray images after sampling
Normal	3270	2756
COVID-19	1275	3267
Other Pneumonia	3000	1356

3.5. Classification based on machine learning techniques

Classification is a predictive modeling technique in which a class label is predicted based on the categories of the input data. For the classification of chest X-ray images, this study involved three machine learning classifiers, namely Support Vector Machine, k-nearest neighbors, and random forest classifiers, for the classification of COVID-19, pneumonia, and normal images.

3.5.1. Support Vector Machine (SVM)

SVMs are machine learning algorithms used for classification and regression. VM is another popular supervised learning method. The SVM classifier creates a model that allocates new data points to one of the defined categories. SVM can be used for various tasks, including linear classification. SVMs are capable of performing nonlinear classification with high efficiency by using a kernel technique. The following kernels are frequently used in SVM: linear, polynomial, radial basis function (RBF), and sigmoid kernels. Compared to other classification algorithms, SVM has substantial advantages in terms of speed, efficiency, and accuracy.

3.5.2. K-Nearest Neighbors (K-NN)

The K-NN model is a classification and regression nonparametric approach. A nonparametric approach indicates that no assumptions are made about data distribution. Here, K is the number of nearest neighbors. It helps to determine most class. It simply calculates the distance between a sample data point and all other training data points. The number of neighbors (K) in the K-NN model is a hyperparameter that must be selected at the time of model prediction. The most important challenge in K-NN is determining the appropriate value of K. A low value of K may result in overfitting, which occurs when the model performs well during the training phase but poorly during the testing phase. A large value of K makes it computationally time-consuming to build the K-NN model. To optimize the results, the K-NN algorithm was tested with different possible values of K ranging from 1 to 10. The model with the highest accuracy can be considered the best option.

3.5.3. Random Forest Classifier (RFC)

The RFC is a Supervised Machine Learning Algorithm used in Classification and Regression problems. It constructs decision trees from various samples and uses the majority vote for classification and the average for regression. All trees in the forest must make predictions for the same input. As a result, it is a time-consuming procedure.

3.6. Performance Evaluation Metrics Parameters

The classifier with better accuracy was determined based on the classifier performance analysis and evaluation. A confusion matrix was used to compare the predicted values from the model with actual values. We computed performance metrics, including sensitivity, f1-score, Classification Error (CE), accuracy, precision, Mat-thews corr-elation coeffi-cient (MCC), and classifi-cation succ-ess in-dex (CSI), using the terms true positive, false positive,false negative, and true negative. The definition and interpretation of each metric are described in detail. In this study, we tackle a multi-classification challenge, which necessitates model assessment via a multi-class confusion matrix [44].

A confusion matrix is a table used to describe the classification algorithm performance. Visualize and summarize the performance of classification algorithms. It summarizes correct and incorrect predictions broken into categories. The confusion matrix comprises four main parameters that are used to create the classifiers measuring metrics. These four outcomes are described below:

- True Positive (TP): means that the actual and predicted values are the same.
- True Negative (TN): This represents the number of predictions that the classifier correctly predicted that the negative class would be negative.
- False Positive (FP): negative class predicts a positive category.
- False Negative (FN): positive cases were misclassified to other classes.

Accuracy describes the accuracy of the model. Precision how precise or accurate the prediction of your model is. Recall how sensitive your model is. The model correctly classified positive values. Classifier correct predictions of negative samples out of all negative samples represent specificity. F1-score is the harmonic mean of precision and recall values. The Matthew correlation coefficient (MCC) [45] is a composite estimate index that integrates sensitivity and specificity. As a correlation coefficient, it occurs in the range of 1 to +1. Perfect prediction is an indicator of +1, average random prediction is indicated by 0, and –1 is an inverse prediction. The classification success index (CSI) [46] is a tool for evaluating the efficiency of a classification model by determining the percentage of correctly classified samples out of all samples. Classification Error (CE): The fraction of predictions that were incorrect. It is also known as Misclassification Rate. The formulas for accuracy, precision, recall, specificity, f1 score, MCC, CSI, and CE are presented in equations (6)–(13), as follows:

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

$$\text{Precision} = \frac{\mathbf{TP}}{\mathbf{TP+FP}} \quad 7)$$

$$\text{Recall} = \frac{\mathbf{TP}}{\mathbf{TP+FN}} \quad 8)$$

$$\text{Specificity} = \frac{\mathbf{TN}}{\mathbf{TN+FP}} \quad 9)$$

$$\text{F1-Score} = \mathbf{2} * \frac{\mathbf{Precision*Recall}}{\mathbf{Precision+Recall}} \quad 10)$$

$$MCC = \frac{(\boldsymbol{TP \star TN}) - (\boldsymbol{FP * FN})}{\sqrt{(\boldsymbol{TP+FP}) * (\boldsymbol{TP+FN}) * (\boldsymbol{TN+FP}) * (\boldsymbol{TN+FN})}} \quad (11)$$

$$CSI = TP/(TP + FP + FN) \quad 12)$$

$$CE = \frac{\mathbf{FP+FN}}{\mathbf{TP+TN+FP+FN}} \quad 13)$$

4. Experimental results and discussion

The results obtained by the proposed multiclass model are discussed. Python programing language was used for the implementation of the X-ray image classification system. The experiments were performed in the Google Colaboratory. A 12-GB NVIDIA Tesla K80 GPU was provided by it, and it could be used for a maximum of 12 h. The experiment was conducted using X-ray images of three categories: normal, COVID-19, and other pneumonia. To solve the problem of imbalance in medical datasets, we applied a hybrid sampling algorithm combining (SMOTE) and (ENN) with three different supervised learning techniques on datasets after extracting the images' features using an oriented gradient histogram algorithm and measuring their performances in terms of accuracy, precision, recall, specificity, f1-score, MCC, CSI, and CE.

The experiment with the k-NN algorithm can be tested using different k values ranging from one to ten. The best results were obtained with k = seven. From Figure 8, a plot was derived between the error rate and K denoting values. Using an error plot, we can see that the minimum error rate was 0.013 at k = 7.

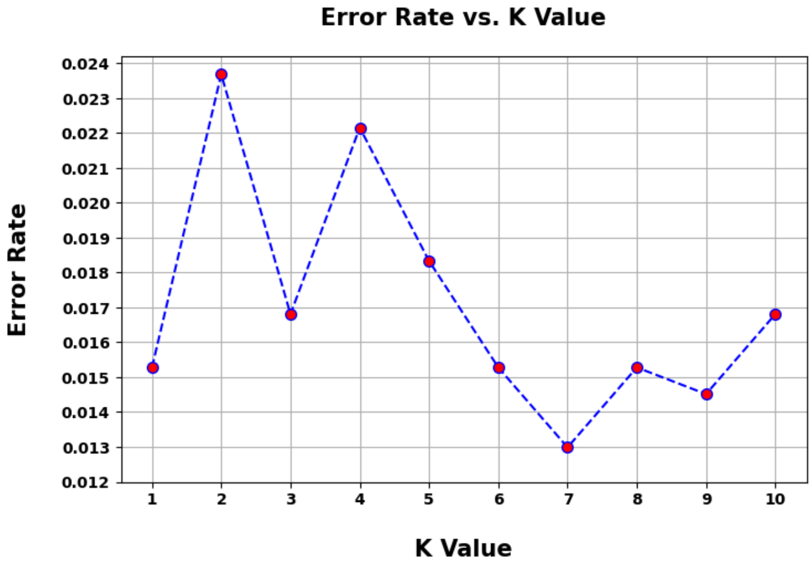


Figure 8. Error rate vs k value graph for k-NN.

As shown in Figure 9, the optimal value of k was seven because the test accuracy achieved its maximum value. We obtained an accuracy of 0.987 at k = 7. Because we already derived the error plot and obtained the minimum error at k = 7, we obtained better efficiency at that k value.

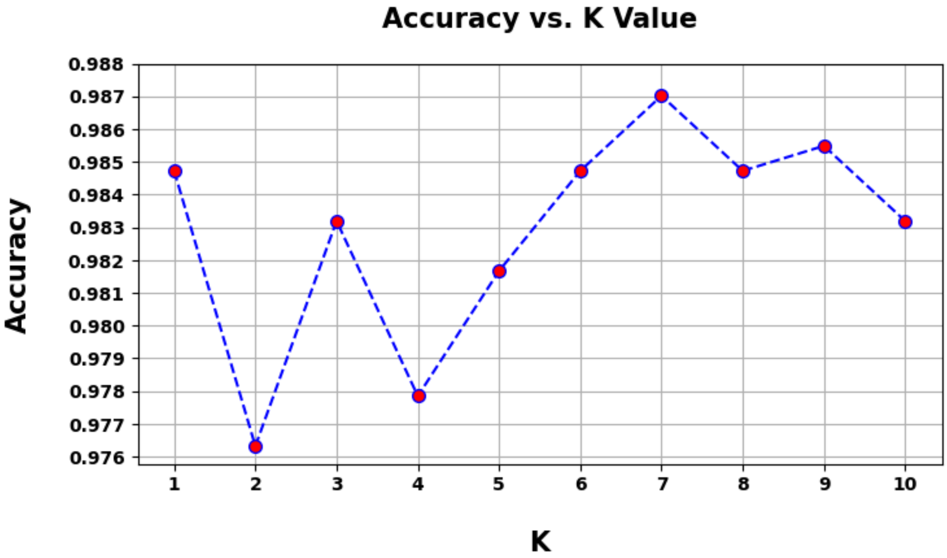


Figure 9. It is feasible to deduce from this graph that the optimal value of K is equal to seven.

The efficacy of the proposed system was ascertained by constructing a confusion matrix by observing the accurate and inaccurate predictions made by the model. The confusion matrices are presented in Figure 10-12 for the three machine learning techniques; SVM, k-NN, and RFC, acquired during experiments on the testing dataset. In the matrix provided, the target and predicted classes are shown on the Y-axis and X-axis.

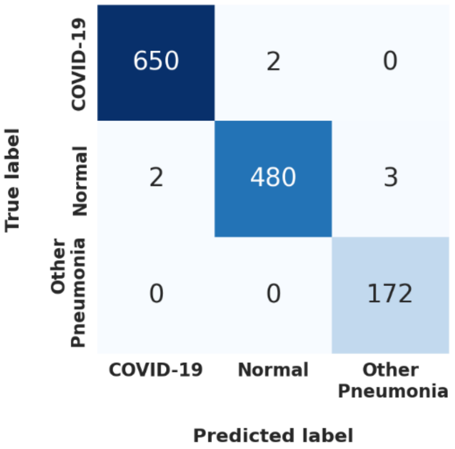


Figure 10. Confusion matrix results of SVM with SMOTE-ENN results in a testing dataset.

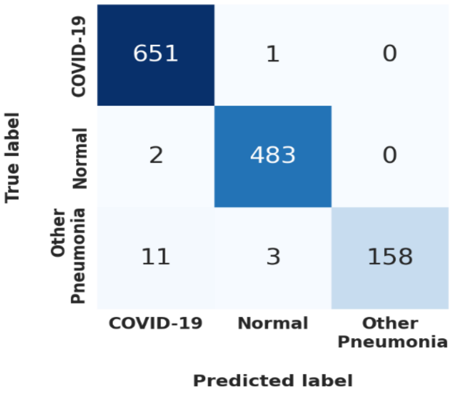


Figure 11. Confusion matrix results of k-NN with SMOTE-ENN results in a testing dataset.

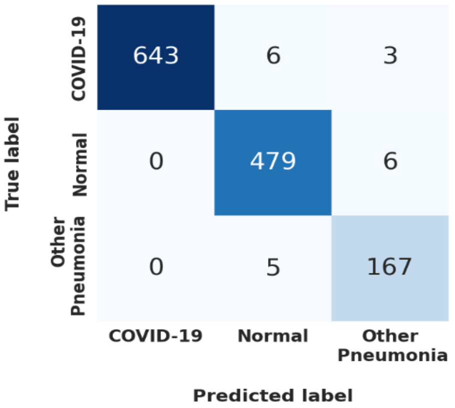


Figure 12. Confusion matrix results of RFC with SMOTE-ENN results in a testing dataset.

Table 2 presents a comparison between three different supervised learning techniques. The evaluation of the three techniques was undertaken to help with the selection of a fitting model for image classification, bearing in mind the test results of these techniques. Among these techniques, SVM demonstrated the highest accuracy of 99.47%. Moreover, SVM surpassed some evaluation metrics, including precision, recall, specificity, f 1-score, MCC, CSI, and CE, achieving percentages of 99.19%99.55%99.73%99.37%99.11%98.75% and 0.36,.The k-NN classifier obtained an accuracy of 98.70%, moreover, in conjunction with alternative metrics for evaluation such as prec-ision, reca-ll, speci-ficity, F1-score, MCC, CSI, and CE, of 99.07%, 97.10%, 99.18%, 98.08% 97.83%, 96.18%, and 0.87, respectively. However, RFC demonstrated the worst performance by averaging an accuracy rate of 98.47%, an accuracy rate of 97.55%, a recall accuracy rate of 98.16%, specificity of 99.29%, an f1_score of 97.85%, a Mathew correlation coefficient of 97.46%, and a correct classification rate of 95.82% and 1.02% CE.

Table 2. "The accuracy, precision, recall, specificity, F1-score, MCC, CSI, and CE for the proposed approach and three classifiers".

Metrics (%)	SVM	k-NN	RFC
Accuracy	99.47	98.70	98.47
Precision	99.19	99.07	97.55
Recall	99.55	97.10	98.16
Specificity	99.73	99.18	99.29
F1-score	99.37	98.08	97.85
Matthew correlation coefficient	99.11	97.83	97.46
Classification success index	98.75	96.18	95.82
Classification Error	0.36	0.87	1.02

The proposed approach should be compared to the results of the literature review. The results of the proposed solution are presented in Table 3.

Table 3. "A comparison of the outcomes between the proposed system and methodologies found in the literature".

Approach / Refs. / Author	Dataset classes	Number of images in dataset	Accuracy
Montalbo [32]	Normal, Covid-19 and Pneumonia	9208	97.41%
Nagi et al. [30]	Covid-19, lung opacity and healthy	19820	94.21%
Haque et al. [33]	Normal, Covid-19 and Pneumonia	5863	97.56%
Mahmoud, et al. [30]	Normal, Covid-19 and Pneumonia	6310	>95%
Proposed Method	Covid-19, normal and pneumonia	7545	99.47%

5. Conclusions

COVID-19 and pneumonia are the most fatal lung diseases. In this paper, Lung disorders can be classified using machine learning models. In the proposed methodology, a successful image enhancement using CLAHE. In addition, a histogram of the oriented gradient algorithm is used to extract the image features. We applied three supervised machine learning classifiers (SVM, KNN and RFC) to classify pneumonia, COVID-19, and normal lung CXRs. An imbalance dataset is one whose number of items in one category is significantly greater than or less than the number in another category. Therefore, the target variable distribution does not spread just as evenly across different categories: With any learner, any learned theorem may be biased; when a model is evaluated on out-of-sample data, this also presents problems. The implementation of SMOTE-ENN hybrid sampling can help improve the accuracy of the model. The proposed method performs multi-class classification with 99.47% accuracy. In future studies our goal is to interrogate various datasets under multiple label conditions. Using the proposed techniques, we want to test how they fit into other diseases. We also want to see if they can prove effective for multiple chest diseases. Future researchers should use an ensemble learning method to improve the performance of the classification method.

References

1. A. Bhujel, N.-E. Kim, E. Arulmozhi, J. K. Basak, and H.-T. Kim, "A Lightweight Attention-Based Convolutional Neural Networks for Tomato Leaf Disease Classification," Agriculture, vol. 12, no. 2, p. 228, Feb. 2022, doi: 10.3390/agriculture12020228. [Online]. Available: http://dx.doi.org/10.3390/agriculture12020228

2. E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7. [Online]. Available: http://dx.doi.org/10.1038/s41591-018-0300-7
3. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x. [Online]. Available: http://dx.doi.org/10.1038/s42256-019-0048-x
4. "2. Classification and Diagnosis of Diabetes:Standards of Medical Care in Diabetes—2021," Diabetes Care, vol. 44, no. Supplement_1, pp. S15–S33, Dec. 2020, doi: 10.2337/dc21-s002. Available: https://doi.org/10.2337/dc21-s002
5. F. Mostafa, L. A. Elrefaei, M. M. Fouda, and A. Hossam, "A survey on AI Techniques for Thoracic Diseases diagnosis Using Medical Images," Diagnostics, vol. 12, no. 12, p. 3034, Dec. 2022, doi: 10.3390/diagnostics12123034. Available: https://doi.org/10.3390/diagnostics12123034
6. A. Bernheim et al., "Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection," Radiology, vol. 295, no. 3, p. 200463, Jun. 2020, doi: 10.1148/radiol.2020200463. Available: https://doi.org/10.1148/radiol.2020200463
7. G. Lv, J. Yuan, X. Xiong, and M. Li, "Mortality rate and characteristics of deaths following COVID-19 vaccination," Frontiers in Medicine, vol. 8, May 2021, doi: 10.3389/fmed.2021.670370. Available: https://doi.org/10.3389/fmed.2021.670370
8. Q. Li et al., "Early transmission dynamics in Wuhan, China, of novel Coronavirus–Infected pneumonia," New England Journal of Medicine/˜the œNew England Journal of Medicine, vol. 382, no. 13, pp. 1199–1207, Mar. 2020, doi: 10.1056/nejmoa2001316. Available: https://doi.org/10.1056/nejmoa2001316
9. J. A. Edlow and P. J. Pronovost, "Misdiagnosis in the emergency department," JAMA, vol. 329, no. 8, p. 631, Feb. 2023, doi: 10.1001/jama.2023.0577. Available: https://doi.org/10.1001/jama.2023.0577
10. T. Vos et al., "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," Lancet, vol. 396, no. 10258, pp. 1204–1222, Oct. 2020, doi: 10.1016/s0140-6736(20)30925-9. Available: https://doi.org/10.1016/s0140-6736(20)30925-9
11. C. D. Chen et al., "Comparing Tau PET Visual Interpretation with Tau PET Quantification, Cerebrospinal Fluid Biomarkers, and Longitudinal Clinical Assessment," Journal of Alzheimer's Disease, vol. 93, no. 2, pp. 765–777, May 2023, doi: 10.3233/jad-230032. Available: https://doi.org/10.3233/jad-230032
12. M. Negahnaz et al., "Hidden lesions: a case of burnt remains," Forensic Sciences Research, vol. 8, no. 2, pp. 163–169, Jun. 2023, doi: 10.1093/fsr/owad019. Available: https://doi.org/10.1093/fsr/owad019
13. S. M. Levine and D. Marciniuk, "Global impact of respiratory Disease," Chest, vol. 161, no. 5, pp. 1153–1154, May 2022, doi: 10.1016/j.chest.2022.01.014. Available: https://doi.org/10.1016/j.chest.2022.01.014
14. Z. Wu and J. M. McGoogan, "Characteristics of and important lessons from the coronavirus Disease 2019 (COVID-19) outbreak in China," JAMA, vol. 323, no. 13, p. 1239, Apr. 2020, doi: 10.1001/jama.2020.2648. Available: https://doi.org/10.1001/jama.2020.2648
15. A. A. Nasser and M. A. Akhloufi, "Deep learning methods for chest disease detection using radiography images," *SN Computer Science/SN Computer Science*, vol. 4, no. 4, May 2023, doi: 10.1007/s42979-023-01818-w. Available: https://doi.org/10.1007/s42979-023-01818-w
16. Y. Qiu, F. H. Lin, W. Chen, and M. Xu, "Pre-training in medical Data: a survey," *Deleted Journal*, vol. 20, no. 2, pp. 147–179, Feb. 2023, doi: 10.1007/s11633-022-1382-8. Available: https://doi.org/10.1007/s11633-022-1382-8
17. C. Tapé, K. Byrd, S. Aung, J. R. Lonks, T. P. Flanigan, and N. Rybak, "COVID-19 in a Patient Presenting with Syncope and a Normal Chest X-ray.," *Rhode Island Medical Journal*, vol. 103, no. 3, pp. 50–51, Mar. 2020, Available: https://pubmed.ncbi.nlm.nih.gov/32226962/
18. D. Zhao *et al.*, "A comparative study on the clinical features of coronavirus 2019 (COVID-19) pneumonia with other pneumonias," *Clinical Infectious Diseases/Clinical Infectious Diseases (Online. University of Chicago. Press)*, vol. 71, no. 15, pp. 756–761, Mar. 2020, doi: 10.1093/cid/ciaa247. Available: https://doi.org/10.1093/cid/ciaa247
19. "Health and economy." Available: https://eurohealthobservatory.who.int/themes/observatory-programmes/health-and-economy
20. M. Hoffman MD, "Lung diseases overview," *WebMD*, Dec. 03, 2022. Available: https://www.webmd.com/lung/lung-diseases-overview
21. World Health Organization: WHO, "Pneumonia in children," Nov. 11, 2022. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia
22. "COVID-19 cases | WHO COVID-19 dashboard," Datadot. Available: https://covid19.who.int/
23. Д. М. Шамаев, "Synthetic datasets and medical artificial intelligence specifics," in *Lecture notes in networks and systems*, 2023, pp. 519–528. doi: 10.1007/978-3-031-21438-7_41. Available: https://doi.org/10.1007/978-3-031-21438-7_41
24. P. P. Wagle and M. Kumar, "A comprehensive review on the issue of class imbalance in predictive modelling," in *Lecture notes in electrical engineering*, 2022, pp. 557–576. doi: 10.1007/978-981-19-5482-5_48. Available: https://doi.org/10.1007/978-981-19-5482-5_48
25. H. Shi, C. Wu, B. Tao, J. Chen, Y. Li, and H. Wu, "Identify essential genes based on clustering based synthetic minority oversampling technique," *Computers in Biology and Medicine*, vol. 153, p. 106523, Feb. 2023, doi: 10.1016/j.compbiomed.2022.106523. Available: https://doi.org/10.1016/j.compbiomed.2022.106523

26. S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0276-2. Available: https://doi.org/10.1186/s40537-019-0276-2
27. O. Stephen, M. Sain, U. J. Maduh, and D. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," *Journal of Healthcare Engineering*, vol. 2019, pp. 1–7, Mar. 2019, doi: 10.1155/2019/4180949. Available: https://doi.org/10.1155/2019/4180949
28. R. Zagrouba *et al.*, "Modelling and simulation of COVID-19 outbreak prediction using supervised machine learning," *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, vol. 66, no. 3, pp. 2397–2407, Jan. 2021, doi: 10.32604/cmc.2021.014042. Available: https://doi.org/10.32604/cmc.2021.014042
29. A. Rahman *et al.*, "Supervised Machine Learning-Based Prediction of COVID-19," *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, vol. 69, no. 1, pp. 21–34, Jan. 2021, doi: 10.32604/cmc.2021.013453. Available: https://doi.org/10.32604/cmc.2021.013453
30. A. T. Nagi, M. J. Awan, M. A. Mohammed, A. Mahmoud, A. Majumdar, and O. Thinnukool, "Performance analysis for COVID-19 diagnosis using custom and State-of-the-Art deep Learning models," *Applied Sciences*, vol. 12, no. 13, p. 6364, Jun. 2022, doi: 10.3390/app12136364. Available: https://doi.org/10.3390/app12136364
31. A. Z. Khuzani, M. Heidari, and S. Shariati, "COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images," *Scientific Reports*, vol. 11, no. 1, May 2021, doi: 10.1038/s41598-021-88807-2. Available: https://doi.org/10.1038/s41598-021-88807-2.
32. F. J. P. Montalbo, "Truncating fined-tuned vision-based models to lightweight deployable diagnostic tools for SARS-CoV-2 infected chest X-rays and CT-scans," *Multimedia Tools and Applications*, vol. 81, no. 12, pp. 16411–16439, Mar. 2022, doi: 10.1007/s11042-022-12484-0. Available: https://doi.org/10.1007/s11042-022-12484-0
33. K. F. Haque, F. F. Haque, L. Gandy, and A. Abdelgawad, "Automatic Detection of COVID-19 from Chest X-ray Images with Convolutional Neural Networks," *International Conference on Computing, Electronics & Communications Engineering*, Aug. 2020, doi: 10.1109/iccece49321.2020.9231235. Available: https://doi.org/10.1109/iccece49321.2020.9231235
34. L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Scientific Reports*, vol. 10, no. 1, Nov. 2020, doi: 10.1038/s41598-020-76550-z. Available: https://doi.org/10.1038/s41598-020-76550-z
35. M. Ragab, S. Alshehri, N. A. Alhakamy, R. F. Mansour, and D. Koundal, "Multiclass classification of chest X-Ray images for the prediction of COVID-19 using capsule network," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–8, May 2022, doi: 10.1155/2022/6185013. Available: https://doi.org/10.1155/2022/6185013
36. "Curated Chest X-Ray Image dataset for COVID-19," *Kaggle*, Nov. 09, 2020. Available: https://www.kaggle.com/datasets/unaissait/curated-chest-xray-image-dataset-for-covid19
37. S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, Sep. 1987, doi: 10.1016/s0734-189x(87)80186-x. Available: https://doi.org/10.1016/s0734-189x(87)80186-x
38. M. Siddhartha and A. Santra, "COVIDLite: A depth-wise separable deep neural network with white balance and CLAHE for detection of COVID-19," *arXiv.org*, Jun. 19, 2020. Available: https://arxiv.org/abs/2006.13873
39. GeeksforGeeks, "Data Pre-Processing with Sklearn using Standard and Minmax scaler," *GeeksforGeeks*, Feb. 03, 2022. Available: https://www.geeksforgeeks.org/data-pre-processing-wit-sklearn-using-standard-and-minmax-scaler/
40. "sklearn.preprocessing.StandardScaler," *Scikit-learn*. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
41. N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Jul. 2005, doi: 10.1109/cvpr.2005.177. Available: https://doi.org/10.1109/cvpr.2005.177
42. M. Kuhn and K. Johnson, *Applied Predictive Modeling*. 2013. doi: 10.1007/978-1-4614-6849-3. Available: https://doi.org/10.1007/978-1-4614-6849-3
43. G. E. a. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explorations*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735. Available: https://doi.org/10.1145/1007730.1007735
44. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002. Available: https://doi.org/10.1016/j.ipm.2009.03.002
45. K. R. Chaudhuri, "The Parkinson's disease sleep scale: a new instrument for assessing sleep and nocturnal disability in Parkinson's disease," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 73, no. 6, pp. 629–635, Dec. 2002, doi: 10.1136/jnnp.73.6.629. Available: https://doi.org/10.1136/jnnp.73.6.629
46. S. E. Nassar, I. Yasser, H. M. Amer, and M. A. Mohamed, "A robust MRI-based brain tumor classification via a hybrid deep learning technique," *˜the œJournal of Supercomputing/Journal of Supercomputing*, vol. 80, no. 2, pp. 2403–2427, Aug. 2023, doi: 10.1007/s11227-023-05549-w. Available: https://doi.org/10.1007/s11227-023-05549-w