ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 10, No. 6, December 2022, pp.168-180

Enhanced feature selection algorithm for pneumonia detection

Salma Hameedi Abdullah [1], Wafaa M. Salih Abedi [2], Raghad Mohammed Hadi [3]
[1]University of Technology, Computer Engineering Department, Baghdad, Iraq
[2]City University Ajman, Ajman, UAE
[3]Al-Mustansiriy University, Medicine College, Baghdad, Iraq

ABSTRACT

Pneumonia is a type of lung disease that can be detected using X-ray images. The analysis of chest X-ray images is an active research area in medical image analysis and computer-aided radiology. This research aims to improve the accuracy and efficiency of radiologists' work by providing a technique for identifying and categorizing diseases. More attention should be given to applying machine learning approaches to develop a robust chest X-ray image classification method. The typical method for detecting Pneumonia is through chest X-ray images but analyzing these images can be complex and requires the expertise of a radiographer. This paper demonstrates the feasibility of detecting the disease using chest X-ray images as datasets and a Support Vector Machine combined with a Naive Bayesian classifier, with PCA and GA as feature selection methods. The selected features are essential for training many classifiers. The proposed system achieved an accuracy of 92.26%, using 91% of the principal component. The study's result suggests that using PCA and GA for feature selection in chest X-ray image classification can achieve a good accuracy of 97.44%. Further research is needed to explore the use of other data mining models and care components to improve the accuracy and effectiveness of the system.

Keywords: Feature Selection Algorithm; Support Vector Machine (SVM); Genetic Algorithm (GA); Naïve Bayesian; and Principal Component Analysis (PCA); Pneumonia.

Corresponding Author:

Salma Hameedi Abdullah
Computer Engineering Department,
University of Technology, Baghdad, Iraq
Email: Salma.H.Abdullah@uotechnology.edu.iq

1. Introduction

Infectious diseases such as Pneumonia can be a significant threat to human health. Pneumonia is a type of lung infection that can be diagnosed using X-ray images. Classification of chest radiographs is an active area of research in medical image analysis and computer-aided radiology analysis. Researchers face several challenges, including relatively low specificity of pathogen detection and incomplete quantification to assess the severity and predict outcomes [1]. Due to the similar appearance of imaging of infectious and inflammatory diseases, there is a low imaging specificity [2]. The powerful tool for diagnosing and managing many medical conditions is medical imaging. Types of medical images include radiography, computed tomography, and fluoroscopy. X-ray imaging is a valuable radiography type, and its analysis is performed traditionally by human experts and therefore requires significant human capital [3]. In developing and underdeveloped countries, Pneumonia is the leading cause of child deaths under the age of five worldwide, where risk issues like high levels of pollution, poor living conditions, and overcrowding are prevalent. Radiological studies (usually x-ray scans) are used to diagnose Pneumonia but are subject to subjective variability and can vary between radiologists. Therefore, need for additional reliable methods to identify signs of Pneumonia from X-rays [4].

Acquiring the image, creating the image, image analysis, and image-based reporting are all part of the processing and interpretation of medical images. Medical image analysis has evolved into various fields, including pattern recognition, image mining, computer visualization, and machine learning. Recently, computer-aided diagnosis has been a part of the advancement of medical image analysis to provide automatic detection and classification of various diseases [5].

Chest X-rays can appear normal or show signs of illness such as lung cancer or Pneumonia to a trained radiologist. Pneumonia is the most common infectious lung disease, a lung infection caused by bacteria, viruses, or fungi [6]. Pneumonia can be dangerous for babies, older adults, and patients on ventilators in hospitals. Additionally, it is a severe disease in developing countries where many people are impoverished and lack access to health care. In an image, features represent information [7], and to recognize the information needs to be viewed from multiple perspectives and analyzed to extract the vital information in a specific domain such as satellite images, medical images, etc., known as feature selection [8].

Image information represents the features or characteristics of the image. Feature or characteristic mining plays an essential role in analyzing images. Principal Component Analysis (PCA) is a linear technique that converts a set of connected characteristics into unconnected characteristics called principal components using an orthogonal transformation. And the information aligns in axes, where the tremendous variety of principal components arise first. Kernel methods calculate principal components through feature space nonlinear mapping and are finally located in a nonlinearly distorted space [9]. The data mining techniques are used to evaluate the performance of comparing and analyzing mined features on linear and nonlinear feature space, such as linear regression, which is a technique to discover the relationship between a series of independent and dependent variables by a suitable equation [10].

Clustering or grouping can be defined as organizing a set of data into groups or classes such that items within a cluster or group are similar to one another but different from the items in other clusters [11, 12].

In this work, we were building a novel feature extraction model that can detect pneumonia disease by using a chest X-ray dataset and applying two classifiers, such as a linear Support Vector Machine combined with a Naive Bayesian classifier, with PCA and GA as feature selection methods.

In this paper, related work is presented in Section 2. In Section 3, the data mining methods are described. The proposed system is explained in Section 4, and Section 5 presents the evaluation performance, experiments, and results. Section 6 discusses future challenges, and the final section presents the conclusion.

1.1. Related work

In this section, we present some of the related work on pneumonia detection. In [13] this study, a CNN method was trained from scratch to detect Pneumonia with 85% accuracy using chest X-ray datasets. The focus is on disease annotation using detailed image features. While the dataset is large and diverse, the quality of the images is still being determined because it is composed of multiple sources. By detailing the process used for extracting features to classify the data, the study provides insights into understanding the data that have not been explored in other works. This leads to a reproducible approach that readers can follow in the preprocessing stage of the study.

In [14], this work developed CNN model VGG16 and Xception for fine-tuning, substantially altered Xception design with two fully linked levels add-on and SoftMax activation mechanism with multiple-output tiers to reduce noise and improve alignment from different diseased regions. It also incorporates global branches to minimize the influence of local branches on the lost discriminatory features. Using the chestXray-14 dataset, the approach captured different CNN features and achieved an accuracy of 0.87. However, this approach is constrained in terms of parameter changes and is resistant to any changes that would prevent the model from predicting a range of data. To detect and diagnose Pneumonia, the experiment used a 121-layer CNN based on the CheXNet algorithm and chest X-ray images as inputs.

In [15], the study utilizes a dilated convolutional CNN method with residual junctions to detect Pneumonia. The method used 49 convolutional layers with two dense layers and achieved 90.05% test accuracy. The method consumed time because it used a substantial convolutional layer and focused on the need for accurate and efficient medical diagnosis. Researchers worldwide are using deep learning methods and machine learning to achieve this goal. And a feed-forward neural network for the detection of tuberculosis diseases where proposed.

The authors in [16], proposed an approach based on machine learning techniques for early and cost-effective detection of Pneumonia. They designed a method to detect the presence of Pneumonia and classify it as microbial or viral based on examining chest radiographs. They achieved a three-class classification based on features that capture different aspects of the images. To balance the sample size in the dataset, they applied an oversampling method and extracted geometric and global features from the chest X-ray images using a deep learning architecture.

In [17], a predictive model based on an SVM with RBF kernel was developed to predict the readmission rate of hospitalized pneumonia patients after discharge from the hospital. The model achieved 83.85% and 82.24% accuracy and was a useful tool for identifying high-risk pneumonia patients.

In [18], the study proposed a convolution neural network model to reduce the deaths caused by Pneumonia. The proposed model achieved 85.6% and 92.31% accuracy, respectively, which is an improvement over previous models. One limitation of the methods is the need for more spatial invariance. This study used transfer learning and deep learning to achieve higher accuracy potentially.

In [19], the authors proposed a framework for forecasting lung diseases, using disease prediction by feature extraction and Region of Interest (ROI) detection. The authors used two public chest X-ray image datasets for acquisition. The authors applied median filtering and histogram equalization to improve image quality, which is often degraded in X-ray images.

In [20], the work presents a computer-assisted diagnosis of Pneumonia using collaborative learning to streamline the analysis process on chest X-ray images. The approach is based on Convolutional Neural Network (CNN) models, and the results are found by merging the mined features through the testing stage. The method achieves an accuracy of 93.91% and an F1-score of 93.88% in the testing stage.

In [21], the study classifies musculoskeletal radiographs and bone X-ray images into no break. It breaks groups using four different classifiers, decision trees, Logistic Regression, linear SVM, and LBF SVM (Radial Basis Function support vector machine). And the performance evaluation of the classification was performed using five mathematical measures: F1 Score, Accuracy, Specificity, Sensitivity, and Precision, which showed significant improvement.

2. Background

This study uses several data mining methods: SVM for classification and Naïve Bayesian Classifier for the probabilistic classifier, PCA to feature selection, and GA optimization. This section illustrates the four data mining methods used in the proposed approach.

2.1. Support vector machine

Support Vector Machine (SVM) is a classic machine-learning technique for classification. It was developed to solve high-dimensional problems by discovering the hyperplane that maximally splits the different classes in the feature space. SVM is a nonlinear regression prediction method. Given a set of training data $\{(A_1, B_1)... (A_i, B_i)\}$ in RR^n*R, according to an unknown probability distribution P(A, B), SVM aims to classify all the data correctly. The decision boundary, or hyperplane, is the one that maximally separates the different classes. To classify the data, SVM seeks the hyperplane that satisfies the following conditions:

$$\text{If } B_i = +\text{one}; \quad WA_i + y >= \text{one} \qquad (1)$$

$$\text{If } B_i = -\text{one}; \; WA_i + y <= \text{one} \qquad (2)$$

$$\text{Intended for all i; } B_i(W_i + y) >= \text{one} \qquad (3)$$

In the equations, A is a feature vector, and w is a weight vector. Therefore, the distance must be greater than zero between the hyperplane and the closest training data points (called support vectors) to classify the data. Among all possible hyperplanes, SVM selects the one where the distance between the hyperplane and the closest data points (also called the margin) is maximized [22]. If the training data is linearly separable and each training vector is located in the same class, then the selected hyperplane is the farthest away from the training data. This desired hyperplane also maximally splits the data into two classes. The distance between a data point and the hyperplane can be calculated using the dot product of A and w because A is on the hyperplane [23]. Similarly, for the other data points, we have a similar formula.

Thus, we can determine the summed coldness after unraveling the hyperplane to adjacent opinions by fixating on and deducing the dual detachments. Receiving Determined Boundary by:

$$\text{Boundary} = \text{Mar} = 2 / \|\text{Bo}\| \qquad (4)$$

Exploiting the boundary is similar to the minimum.

$$Q(\text{Bo}) = \tfrac{1}{2}\, Bo' \text{Bo} \qquad (5)$$

For all points $\{(A_i, B_i)\}$: B_i ($Bo^T A_i$ + b) ≥ 1, it is necessary to enhance a quadratic objective function to linear constraints. To solve this problem, we must construct a dual problem in which a Lagrange multiplier La_i is associated with each constraint in the primary problem. The solution involves finding $La_1 \ldots La_N$ such that:

$$Q(La) = \Sigma La_i - ½\Sigma\Sigma\, La_i\, La_j\, b_i \quad b_j\ \ a_i^T \quad (6)$$

$$a_j$$

It was exploited, and we compute $\Sigma La_i\, b_i$ = 0, then $La_i \geq 0$ despite La_i, the solution will have the form: Bo= $\sum La_i\ b_i\ a_i$ and y=b_k-$Bo^T a_k$ for any a_k such that La_k≠0. A piece of non-zero La_i orientations that consistent a_i is a provision vector. Formerly the categorizing purpose determination by the following formula:

$$f(a) = \sum La_i\ \ b_i\, a_i^T\ a + Lb \qquad (7)$$

It relies on an inner product between the input vector A and the support vector a_i. Solving the optimization problem involves calculating the inner product $a_i^T a_j$ between all training vectors. The initial training consists of finding Bo and Lb using the following formula:

Φ(Bo) =½ $Bo^T\ Bo$ is minimalized, and for wholly $(a_i, b_i)\}$, $b_i\ (Bo^T\ a_i\ + Lb)$ greater and equal to one [24, 25].

2.2. Naïve Bayesian classifier

A probabilistic classifier, known as a Naive Bayes classifier, is based on Bayes' theorem and assumes independence. A training set of documents is labeled as A. Each record R is represented by an n-dimensional feature vector [f1, f2... fn], where there are m modules, Cl1, Cl2, ..., Clm. The classifier predicts that record R belongs to the group with the highest posterior probability based on the training set. Therefore, the naïve Bayesian classifier forecasts the record R goes to the group Cli if it achieves the following:

Pro (Cii | Xx), ii Є [1,mm] > Pro(Cjj | Xx),

for 1≤ jj ≤ mm, jj ≠ ii, then

cs(xx)=arg $max_{cs\in C}$ Pro(cs)*Pro(Xx | cs)

Where:

Pro(Ccii | Xx) is the possibility of feature to a reduction in group Ccii

Pro (Ccii) = the previous possibility for group Ccii

Cc(Xx) is determined subsequent cast-off to allocate the group cs having larger pro (Xx|cs). The tutorial-restricted individuality is clarified as this calculation:

Pro(Xx | Cc_{ii}) =Pro($a_1, a_2, \ldots \ldots\ a_{nn}$ |cs) $= \prod_{j=1}^{n}$ pro(a_{jj}| cs_{ii})

Through abridging the intention of Pro(Cc) and Pro(a_{ii}| cs), NB classifier was effortlessly constructed [26].

2.3. Feature selection

It is a required dimensionality reduction method for representing image data, long or short geographical and object-based data, and so on. This model eliminates irrelevant, noisy, and redundant data from the datasets and produces a smaller, more efficient set of features to classify the data better. Feature selection can be performed using two altered methods: wrapper and filter. The filter method does not rely on any learning algorithm and automatically removes features that do not contribute significantly to the performance. The filter method's main advantage is its low computational cost [27].

2.3.1. Genetic algorithm

Genetic algorithms (GA) are a distinguished meta-heuristic method used to solve problems related to optimization. The genetic algorithm can work with a population of possible solutions randomly generated. A small set of genes represents individual solutions. The fitness function determines quality evaluation for each individual (genes). The method involves selecting the first population and continues with three processes: selection, crossover, and mutation. The selection phase includes selecting a set of promising solutions from the initial population. Afterward, pairs of genes are chosen as parents (two randomly selected sets of genes). These parents create two offspring genes using the process described earlier. Crossover is a procedure that includes exchanging genetic information between parent genes. And the final process, offspring genes go through mutation, and the results are potentially improved solutions. The offspring genes are evaluated as described earlier and compared to the genes produced in the earlier generation. The fitness values determine whether the current offspring replaces the previous genes [28].

2.3.2. Principal component analysis

Principal component analysis (PCA) involves finding the main eigenvectors, representing the most spread information along these axes. This aligns all features or dimensions along the axes such that the most significant variance principal component is in the first compared to the other principal components. PCA is considered an essential process for reducing the number of features or dimensions and selecting the most important ones in the data set, thus reducing the computational time of applying classification methods in the proposed system. PCA was applied to the training chest x-ray dataset to identify the most important features or dimensions [29, 30].

3. Description of the proposed system

In this work, a support vector machine is combined with a Naive Bayesian classifier to create a hybrid algorithm that further helps to reduce the computational burden of classifying the Pneumonia features from the dataset into either the Pneumonia class or the normal class. The proposed system consists of three main steps, which are described as follows:

3.1. First step: Preprocessing X-ray images

First step: Preprocessing X-ray images working with x-ray images as the dataset was challenging. Images generated ambiguity needed to clarify whether the machine learning model could accurately classify objects. We use a method for cleaning the dataset before the proposed classifier model training in the proposed system. In the proposed approach, Sobel and Scharr's operators detect edge features to create an image with enhanced boundaries between the objects in the original X-Ray. They are using the resulting images to train a hybrid SVM and Naive Bayesian classifiers to classify the image as normal or abnormal. To enhance the boundaries of the image, the proposed system uses histogram equalization, which spreads the pixel intensity range (0 to 255). As a result, the improved image has an extensive range of intensities and significantly enhanced contrasts and also applies the Sobel/Scharr operator to detect the boundaries within the image. Typically, a 3*3 kernel is used to filter the image horizontally via Hx and vertically via Hy.

$$H_x = \begin{matrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{matrix} * image \qquad H_y = \begin{matrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{matrix} * image$$

The Sobel operator is used to detect the presence of image edges by looking at the difference between the higher and lower intensities of the image or left and right in the horizontal and vertical gradient operations. It shows whether the changes in the image are abrupt or gradual. The Scharr operator, on the other hand, is used to detect edges where the mean squared error is minimized for enhanced clarity and detail. This edge detection phase showed that the structures in chest x-ray images are now more discernible, and the edges will provide the desired visual information for the subsequent phase.

3.2. Second step: Feature selection

Texture descriptors describe 14 pieces of information designed to characterize the image texture. Texture features extraction states the features and the object present in the image. Entropy is calculated using the entropy function, while the gray-level co-occurrence matrix function is used to calculate homogeneity, correlation, energy, contrast, variance, mean, standard deviation, and skewness using the skewness function. In the proposed system, we use PCA and optimize using GA, where PCA is a linear technique used for features extraction method as a result of phase 2 in the proposed system, as shown in the algorithm (1). Therefore, three sets of features, in addition to the complete set of features in the training dataset, will be used in the training of the proposed classifier.

Algorithm 1: Adapted PCA

Input: Proposed chest x-ray learning dataset

Output: Set of PCAs that have associated features and greatest frequent

Steps:

1. Acquire preparation chest x-ray images.
2. Embody all transaction J_i as a direction A_i
3. Calculate the middling transaction:

$$£=\frac{1}{n}\sum_{i=1}^{n} A_i \qquad (8)$$

4. Subtract the mean transaction:

$$Ϙ_i=A_i\text{-}£_i \qquad (9)$$

5. Compute the covariance matric:

$$\mathrm{CC}=\frac{1}{n}Ϙ_m Ϙ_m^T=\mathrm{X}X^T \qquad (10)$$

a. Preliminary CC calculates the eigenvector u_i of $\mathrm{X}X^T$

b. Assume array $\mathrm{X}X^T$ as an N*N array.

6. Compute the eigenvector v_i of $\mathrm{X}X^T$ such that:

$$\begin{aligned} \mathrm{X}X^T v_i \rightarrow \mu_i V_i \rightarrow \mathrm{X}X^T\mathrm{X}V_i \qquad (11) \\ = \mu_i A_{V_i} \rightarrow \mathrm{CC}u_i \\ = \mu_i u_i \end{aligned}$$

where $\mu_i = XV_i$

Compute the µ best eigenvectors of $\mathrm{X}X^T$: $\mu_i = Xv_i$

7. Retain solitary k eigenvectors (l topographies through their standards).

3.3. Third step: Apply classifiers

A training procedure uses a matrix of known contribution data and known data reactions, such as classes, to train a prototype to provide a reliable prediction of how new data will respond. Chest X-ray images with Pneumonia were labeled as type 1, and images without illness were labeled as type 0. The input X-ray images can be represented as a two-dimensional array, where the rows represent individual chest X-ray images (observations) and the columns represent each image's features (predictors). The data array contains a unique row of features selected from different images using PCA and optimized using GA. The hybrid classifiers will be used three times on each of these three groups of features to create the planned classifiers:

1. All features of chest X-ray images.
2. A subgroup of features in chest X-ray images according to the results of applying the PCA and GA techniques. Table 1 presents the indexes of selected features with their descriptions in detail.

Table 1. Datasets description

Sickness	Amount of instance	Amount of features
Covid 19	352	13
Common Fever	179	8
Pneumonia	251	25
Malaria	203	10

The SVM method was used in the initial stage to improve the attribute set provided by extracting features using PCA and optimizing GA. The SVM was used to select attributes and train using the Naive Bayes classifier, storing the trained model in a database. Test data is uploaded, and the trained model from the database is used to classify. The prediction and classification accuracy are calculated. Figure 1 illustrates the proposed system.

Algorithm 2: SVM

Input: Planned chest x-ray training dataset; Proposed chest x-ray testing dataset that needs to be classified.

Output: Chest x-ray testing dataset that classified.

Steps:

1. Reset opinions in learning dataset ultimately as (a_i, b_i) somewhere a is a direction of facts $a_1,\ldots, a_n$ and b is the direction of classes.
2. Reset direction of weightiness W.
3. Allocate all opinions (a, b), then excerpt the hyperplane partition.
4. Uncertainty the hyperplane contributes ideal parting formerly be contingent hyperplane as classifier classical to categorize chest x-ray testing dataset and drive to End otherwise essential do the following steps:
5. Make the best use of the hyper plan by calculation (4) and then for the least consuming calculation (5).
6. Reset multiplier, which is called Lagrange La_i direction $La_1 \ldots\ldots La_n$ consuming calculation (6)
7. Implementation classification using calculation (7)
8. Control the support directions a_i through non-zero La_i.
9. Being contingent, the hyper plan caused decisive support directions as the classifier model to categorize the chest x-ray testing dataset afterward.
10. End.

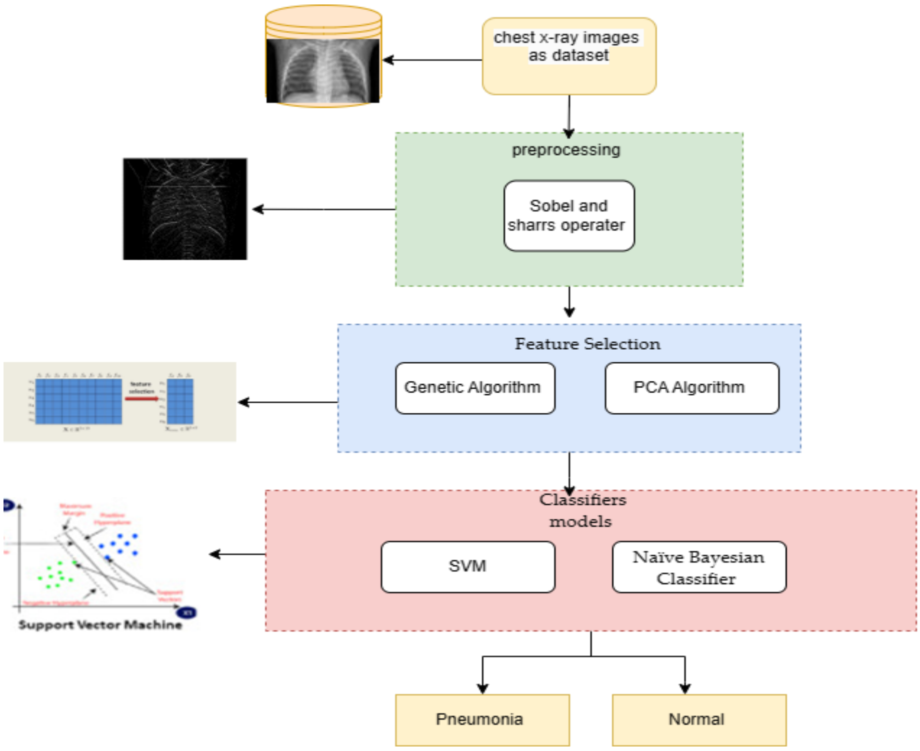

Figure 1. Proposed system diagram

4. Results and discussion

The chest x-ray images were used for training and testing the proposed system; a total of 2500 chest images were used, of which 1300 were pneumonia images, while the rest were standard images. The chest x-ray images were preprocessed to remove unclear features using a Gaussian filter. Then, the proposed system extracted texture and shape-based features using image processing methods. Training of hybrid classifiers on the chest x-ray images was carried out with three groups of selected features: (1) applying a hybrid classifier based on all features, (2) applying a hybrid classifier based on features selected using GA (Genetic Algorithm), and (3) applying a hybrid classifier based on features selected using PCA (Principal Component Analysis).

Consequently, the proposed system has been tested multiple times to evaluate the accuracy of the classifiers. The results of the three trials that produced the best results are presented in this section. Then, the classification models were applied to the chest x-ray image testing dataset to evaluate their performance. The classification results for testing are either TPn (Pneumonia), TNn (normal), false positive (FPo) (not classified as Pneumonia), false negative (FNe) (not classified as normal), or unidentified (novel user activities or novel attacks). Table 2 presents the groups of designated features for Exp1, Exp2, and Exp3.

Table 2. Extracted feature and normalized from one chest x-ray image

No.	Feature	Value	Normalized
1.	Variance	110.012	0.05376985
2.	Mean	110.5	0.06524075
3.	Correlation	89.33	0.05448678
4.	Contrast	17	0.00860318
5.	Entropy	2.0954	0.01075397
6.	skewness	27.23	0.00573545
7.	Homogeneity	96.7	0.04373281
8.	Energy	58.8	0.01935715
9.	Standard deviation	10.544	0.00000000
10.	Zone	1822	0.99366686

To estimate the performance of the proposed collaborative technique for the three datasets, four regular estimate metrics were calibrated: f1-score (F1), recall (Rec), accuracy (Acc), and exactness or precision (Pre). These estimate metrics have the expressions: "True Positive," "False Positive," "True Negative," and "False Negative". Two classes in the dataset are called the "positive" and the "negative" class in the binary classification task. These terms can then be well-defined as follows:

- (TPN) means True Positive, which mentions an example appropriate to the confident class, existence properly categorized through a model.
- (FPO) means False Positive, which denotes an example is appropriate to the undesirable class, existence erroneously categorized as fitting to the positive class.
- (TNN) means True Negative, which mentions an example of going to the undesirable class and correctly classifying via the model.
- (FNE) means False Negative mentions an example going to the confident class, existence erroneously categorized as fitting to the undesirable class.
  The proposed system usage accuracy, sensitivity, specificity, detection rate, and False Alarm Rate present estimate metrics [31-36], which can be labeled as shadows.

$$\text{Accuracy} = \frac{TPn+TNn}{TPn+TNn+FPo+FNe} \quad (12)$$

$$\text{Sensitivity} = \frac{TPn}{TPn+FNn} \quad (13)$$

$$\text{Specificity} = \frac{TNn}{TNn+FPo} \quad (14)$$

$$\text{Detection Rate} = \frac{TPn}{TPn+FNe+unidentified} * 100\% \quad (15)$$

$$\text{False Alarm Rate} = \frac{FP}{TN+FP+unidentified} * 100\% \quad (16)$$

Figure 2 summarizes the values of classification results using all features selected from chest x-ray images. Accuracy is equal to 96.76 %, precision is equal to 94.4%, Specificity is equal to 94.5%, and Sensitivity is equal to 91.06%.

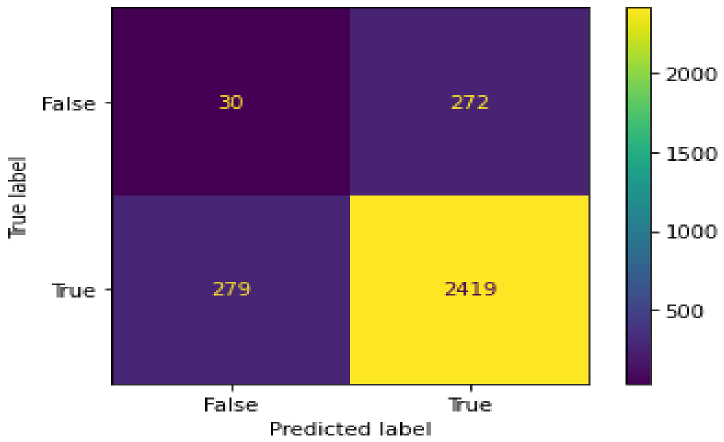

Figure 3: Confusion matrix using all feature extraction

The consequence confusion matrix of consuming GA, a feature in the taxonomy procedure, is exposed in Table 4, anywhere accuracy is equivalent to 97. 44 %, precision is equal to 96.4%, Specificity is equal to 95.35%, and Sensitivity is equal to 97.16%.

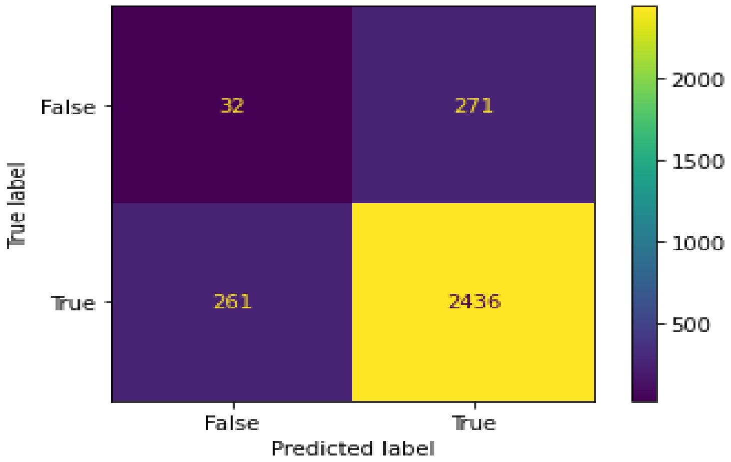

Figure 4. Confusion matrix consuming GA

The consequence confusion matrix of consuming PCA feature in the taxonomy procedure is presented in Table 5, where Accuracy is equivalent to 93.48 %, precision is equal to 94.4%, Specificity is equal to 93.35%, and Sensitivity is equal to 94.16%.

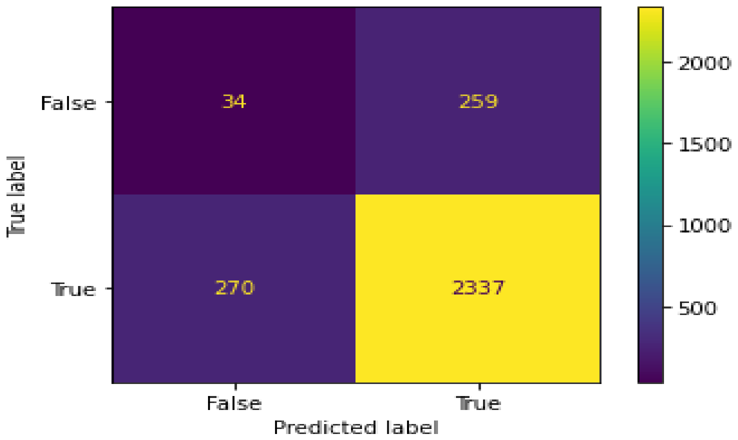

Figure 5. Confusion matrix using PCA

With the offered consequences, cataloging the X-ray images using SVM with PCA provided greater accuracy than using all features and PCA features. In addition, SVM with PCA delivered greater correctness compared to the additional type topographies. The number of topographies used in the taxonomy progression was optional in height, and then the category of the procedure cast off to select these topographies or features is significant to produce greater accuracy. The choice of the finest classical classification would be ended meaningfully, rendering its classification accuracy. The testing process takes step-by-step for this proposed method, enabling users to see the images of chests and X-rays of ordinary and abnormal patients (Figure 2).

Several techniques have been applied to discover thoracic diseases, particularly Pneumonia. While various apparatuses and methods have been used for this purpose, approaches based on machine learning are fairly effective in discovering medical images from image datasets. To make machine learning more effective, it is necessary to have a large and diverse dataset to draw from. Therefore, it is essential to devise a way to collect real-time data to meet the need for a more extensive and diverse dataset. Table 6 compares the proposed system's accuracy with several state-of-the-art methods on chest X-ray datasets.

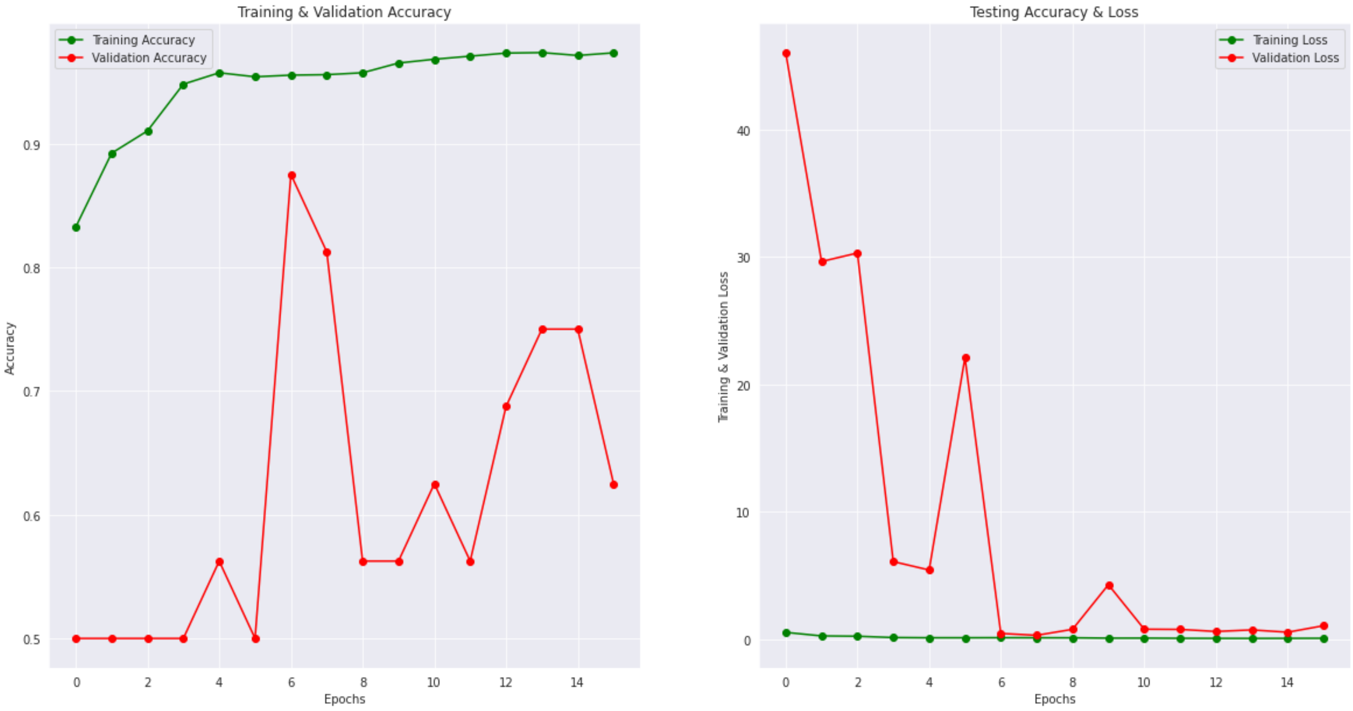

Figure 6. The accuracy of train datasets and test datasets when using all features

Table 3. Comparison of the accuracy of the proposed system with several state-of-the-art methods on chest X-ray datasets.

References	Method	Accuracy%
[13]	CNN models	85.0
[14]	VGG16	87.0
[15]	DeepResidual network	90.05
[18]	SVM through RBF	83.85 and 82.24
[21]	CNN models	93.91
[22]	Machine Learning Paradigm	95.49
Proposed method	PCA, GA, SVM, Naïve Bayesian	97

In the future, there are several problems that we can address using deep learning techniques such as classification, segmentation, and recognition, which are commonly used for clinical diagnoses of various body parts. We can also consider essential terms related to deep learning, such as basic architecture, data augmentation, and feature selection methods. Medical images will continue to be a significant input for deep learning models in the coming years, and deep learning methods are expected to be the primary method for medical image analysis.

5. Conclusion

The pneumonia disease-detecting system was developed to classify and sub-classify Pneumonia. The investigative results show that the proposed method of using PCA and GA for feature selection of chest X-ray images has an excellent accuracy of 97.26 % and can therefore be used for classification with additional methods such as the SVM to improve the accuracy rate and make it more efficient. Combination of algorithms led to an improvement in the proposed system. The system was able to classify the type of Pneumonia with a high accuracy rate of 99% and a low false positive rate of 0.05% using dataset 1, an accuracy rate of 97.44%, and a false positive rate of 2.2% using dataset 2. The system could accurately identify the disease's sub-class in the second level for its modules. In future research, the capabilities of this approach will be explored by incorporating other data mining models and care components in this study.

Declaration of competing interest

The authors declare that they have no known financial or non-financial competing interests in any material discussed in this paper.

Funding information

No funding was received from any financial organization to conduct this research.

References

[1] R. Alsharif, Y. Al-Issa, A. M. Alqudah, I. A. Qasmieh, W. A. Mustafa, and H. Alquran, “PneumoniaNet: Automated detection and classification of pediatric pneumonia using Chest X-ray images and CNN approach,” *Electronics (Basel)*, vol. 10, no. 23, p. 2949, 2021.

[2] E. A. Kim *et al.*, "Viral pneumonia in adults: radiologic and pathologic findings," *Radiographics*, vol. 22 Spec No, no. suppl_1, pp. S137-49, 2002.

[3] W. Jia, *Edge Detection operators for X-ray images based on hessian matrices*. Nova Scotia: University Halifax, 2020.

[4] S. Chattopadhyay, R. Kundu, P. K. Singh, S. Mirjalili, and R. Sarkar, “Pneumonia detection from lung X-ray images using local search aided sine cosine algorithm based deep feature selection method,” *Int. J. Intell. Syst.*, vol. 37, no. 7, pp. 3777–3814, 2022.

[5] M. Ridzuan, A. Bawazir, I. Gollini Navarrete, I. Almakky, and M. Yaqub, “Self-supervision and multi-task learning: Challenges in fine-grained COVID-19 multi-class classification from chest X-rays,” in *Medical Image Understanding and Analysis*, Cham: Springer International Publishing, 2022, pp. 234–250.

[6] F. Faridah, A. Balza, and L. S. Binar, “Lip image feature extraction utilizing Snake’s control points for lip reading applications,” *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 5, no. 4, p. 720, 2015.

[7] A. Ehsani Rad, M. S. Mohd Rahim, and A. Norouzi, “Digital Dental X-Ray Image Segmentation and Feature Extraction,” *TELKOMNIKA Indones. J. Electr. Eng.*, vol. 11, no. 6, 2013.

[8] P. Rajpurkar *et al.*, “CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,” *arXiv [cs.CV]*, 2017.

[9] Roopa and Asha, “Feature extraction of chest X-ray images and analysis using PCA and kPCA,” *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 8, no. 5, p. 3392, 2018.

[10] K. V. Divya, A. Jatti, P. S. Meharaj, and R. Joshi, “Appending active contour model on digital panoramic dental X-rays images for segmentation of maxillofacial region,” in *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2016.

[11] N. A. Hamza, S. H. Jafer, and R. M. Hadi, “3D model retrieval using MeshSIFT descriptor and fuzzy C-means clustering,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 19, no. 3, p. 1452, 2020.

[12] R. Barrientos *et al.*, "Automatic detection of pneumonia analyzing ultrasound digital images," in *2016 IEEE 36th Central American and Panama Convention (CONCAPAN XXXVI)*, 2016.

[13] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," *J. Healthc. Eng.*, vol. 2019, p. 4180949, 2019.

[14] E. Ayan and H. M. Unver, "Diagnosis of pneumonia from chest X-ray images using deep learning," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 2019.

[15] G. Liang and L. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," *Comput. Methods Programs Biomed.*, vol. 187, no. 104964, p. 104964, 2020.

[16] M. Masud *et al.*, "A pneumonia diagnosis scheme based on hybrid features extracted from chest radiographs using an ensemble learning algorithm," *J. Healthc. Eng.*, vol. 2021, p. 8862089, 2021.

[17] B. Gaye, D. Zhang, and A. Wulamu, "Improvement of support vector machine algorithm in big data background," *Math. Probl. Eng.*, vol. 2021, pp. 1–9, 2021.

[18] H. Nezaratian, J. Zahiri, M. F. Peykani, A. Haghiabi, and A. Parsaie, "A genetic algorithm-based support vector machine to estimate the transverse mixing coefficient in streams," *Water Qual. Res. J. Can.*, vol. 56, no. 3, pp. 127–142, 2021.

[19] *Detection and classifcation of lung diseases for Pneumonia and Covid 19 using machine and deep learning techniques*. Journal of Ambient Intelligence and Humanized Computing.

[20] A. Mabrouk, R. P. Díaz Redondo, A. Dahou, M. Abd Elaziz, and M. Kayed, "Pneumonia detection on chest X-ray images using ensemble of deep convolutional neural networks," *Appl. Sci. (Basel)*, vol. 12, no. 13, p. 6448, 2022.

[21] *GLCM-Based Feature Extraction and Medical XRAY Image Classification using Machine Learning Techniques*.

[22] Y. Kong, X. Ma, and C. Wen, "A new method of deep convolutional neural network image classification based on knowledge transfer in small label sample environment," *Sensors (Basel)*, vol. 22, no. 3, p. 898, 2022.

[23] S. Y. Chaganti, I. Nanda, K. R. Pandi, T. G. N. R. S. N. Prudhvith, and N. Kumar, "Image Classification using SVM and CNN," in *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, 2020.

[24] H. A. Owida, A. Al-Ghraibah, and M. Altayeb, "Classification of chest X-ray images using wavelet and MFCC features and Support Vector Machine classifier," *Eng. Technol. Appl. Sci. Res.*, vol. 11, no. 4, pp. 7296–7301, 2021.

[25] K. Almezhghwi, S. Serte, and F. Al-Turjman, "Convolutional neural networks for the classification of chest X-rays in the IoT era," *Multimed. Tools Appl.*, vol. 80, no. 19, pp. 29051–29065, 2021.

[26] N. M. Elshennawy and D. M. Ibrahim, "Deep-pneumonia framework using deep learning models based on chest X-ray images," *Diagnostics (Basel)*, vol. 10, no. 9, p. 649, 2020.

[27] P. Dutta, S. Paul, A. J. Obaid, S. Pal, and K. Mukhopadhyay, "Feature selection based artificial intelligence techniques for the prediction of COVID-like diseases," *J. Phys. Conf. Ser.*, vol. 1963, no. 1, p. 012167, 2021.

[28] M. I. Mahendra and I. Kurniawan, "Optimizing convolutional neural network by using genetic algorithm for COVID-19 detection in a chest X-ray image," in *2021 International Conference on Data Science and Its Applications (ICoDSA)*, 2021.

[29] 29. Bajwa I. S., M. Sh., Asif M. N., Hyder S. I, Ed., *Feature Based Image Classification by using Principal Component Analysis*, vol. 9, no. 2. ICGST-GVIP Journal, 2009.

[30] M. Qasim *et al.*, "PCA-based Advanced Local Octa-Directional Pattern (ALODP-PCA): A texture feature descriptor for image retrieval," *Electronics (Basel)*, vol. 11, no. 2, p. 202, 2022.

[31] J. Q. Kadhim, I. A. Aljazaery, and H. T. H. S. ALRikabi, "Enhancement of online education in engineering college based on mobile wireless communication networks and IoT," *Int. J. Emerg. Technol. Learn.*, vol. 18, no. 01, pp. 176–200, 2023.

[32] H. Th. ALRikabi et al. "Face Patterns Analysis and Recognition System Based on Quantum Neural Network QNN", *International Journal of Interactive Mobile Technologies*, vol.16, no.8, 2022.

[33] I. A. Aljazaery, H. T. S. Alrikabi, and A. Hadi M. Alaidi, "Encryption of color image based on DNA strand and exponential factor," *Int. J. Onl. Eng.*, vol. 18, no. 03, pp. 101–113, 2022.

[34] A.H. M. Alaidi, et al., "Dark Web Illegal Activities Crawling and Classifying Using Data Mining Techniques"', *International Journal of Interactive Mobile Technologies*, vol.16, no.10, p122-139, 2022.

[35] R. Khairy, Directorate General of Education in Babylon, A. Hussein, H. ALRikabi, Directorate General of Education in Babylon, and Wasit University, “The detection of counterfeit banknotes using ensemble learning techniques of AdaBoost and voting,” *Int. J. Intell. Eng. Syst.*, vol. 14, no. 1, pp. 326–339, 2021.

[36] W. M. S. Abedi, “Unconsciousness detection supervision system using faster RCNN architecture,” in *Proceedings of the 2nd International Conference on Future Networks and Distributed Systems*, 2018.