A New Approach Method for Multi Classification of Lung Diseases using X-Ray Images

Sri Heranurweni, Andi Kuniawan Nugroho, Budiani Destyningtias
Electrical Department-Engineering Faculty, Universitas Semarang, Semarang, Indonesia

***Abstract*—Lung disease is one of the most common diseases in today's society. This lung disease's treatment is frequently postponed. This is usually due to a lack of understanding about proper treatment and a lack of clear information about lung disease. Reading the correct X-ray images, which is usually done by experts who are familiar with these X-rays, is one method of detecting lung disease. However, the results of this diagnosis are dependent on the expert's practice schedule and take a long time. This study aims to classify lung disease images using preprocessing, augmentation, and multimachine learning methods, with the goal of achieving high classification performance accuracy with multi-class lung disease. The classification ExtraTrees was obtained from experimental results with unbalanced datasets using a balancing process with augmentation. Precision, Recall, Fi-Score, and Accuracy are 100% for training and testing data 89% for Precision, 88% for Recall, 87 for Fi-Score, and 85% for Accuracy outperform other machine learning models such as Kneighbors, Support Vector Machine (SVM), and Random Forest in classifying lung diseases. The conclusion from this research is that the machine learning approach can detect several lung diseases using X-ray images.**

Keywords*—Augmentation; machine learning; lung disease; prepossessing

I. Introduction

The lungs are one of the organs in the respiratory system that serve as a site for the exchange of oxygen and carbon dioxide in the blood. Polluted air is a common problem, and the air that is inhaled contains many germs that will attack the lungs. Lung disease is a serious disease that affects the human respiratory system and can be fatal if not properly treated. This lung disorder causes sufferers to have difficulty breathing, difficulty performing activities, and a lack of oxygen, which can lead to death if not detected quickly [1]. Tuberculosis, bronchitis, pneumonia, lung cancer, emphysema, and pleuritic are all common lung diseases. It is usually done clinically to detect lung disease/disorders (physical symptoms by a doctor). Aside from clinical examination, X-rays, CT scans, and MRI can be used to diagnose lung disease; however, the latter two methods are more expensive. [2]. Another issue is a lack of public knowledge in reading CT Scan and MRI results, so experts such as doctors or other medical personnel are still required to read them. Many other difficulties, such as complicated backgrounds and the presence of multiple potential abnormalities, make clinical analysis of X-ray images a difficult task. [3]. As a result, manual annotations from experts (radiologists) are required.

Automatic X-ray image analysis is quickly becoming a valuable clinical diagnostic tool. Deep neural networks have recently achieved image classification success and are now widely used for X-ray image classification tasks [4], [5]. Deep learning on chest X-ray images can be used to classify a variety of diseases, including thoracic infections [4], COVID-19 [4]–[7], and lung disorders [8]. Husayn et al. [9] proposed CoroNet, a deep learning model for COVID-19 detection. A deep learning framework has been proposed to detect lung abnormalities in CXR and CT scan images [8]. Using GAN-based synthetic data, Albahli et al. [10] proposed a deep learning model that achieved 87% accuracy and produced results comparable to other techniques. Lung segmentation is another critical task in CXR disease detection. This is extremely useful for determining the severity of tuberculosis [11]. DeTraC's conventional deep neural network architecture is described by Abbas et al. [5].

II. Literature Review

The classification of lung diseases using X-ray images has been explored using various machine learning techniques. A hierarchical classification can be useful in detecting pneumonia due to the disease's hierarchical pattern [12]. Traditional machine learning approaches, such as Support Vector Machine (SVM), k-Nearest Neighbor (KNN), and Decision Tree classifiers, can be used to classify diseases in Chest X-Ray (CXR). They do, however, rely on a mechanism for feature extraction. Convolutional Neural Networks (ConvNets) are another mechanism for feature extraction. Toğacar et al. [13] compared the performance of traditional machine learning models for detecting pneumonia with the Redundancy Maximum Relevance (mRMR) minimum feature selection mechanism. Khatri et al. [14] compared CXR pneumonia images using the Earth Mover's Distance (EMD). Teixeira et al. [15] evaluated and described COVID-19 using lung segmentation. A multimodal approach to disease on CXR can aid in better understanding and elucidation. [16], [17].

Due to a lack of large amounts of annotated data and efficient machine learning algorithms to study some specific features, automatic disease classification on X-ray images is difficult [25]. A variety of data streams and modalities can be used to improve disease prediction accuracy. To train thoracic disease classifiers, text data from diagnostic X-ray images are combined with annotated image data [[17]. A deeply decomposed generative model, as opposed to the traditional approach of directly classifying disease, can be used to generate residual maps for abnormal disease alongside normal images [18]. This method aids in distinguishing between abnormal and normal chest X-ray parts. The semi-supervised

generative model is effective for CXR disease classification [18], [19].

Multiple diseases can still occur at the same time. Single-label classification may be ineffective in this situation, whereas multi-label classification may be effective Albali et al. [20] proposed a CNN-based deep learning method for CXR multi-label classification. In addition, Baltruschat et al. [21] compared several multi-label chest radiograph classification methods based on deep learning. Pathology datasets are rife with class imbalances. Pathology datasets are rife with class imbalances. As a result, the trained classifier is vulnerable to bias towards the majority class. Appropriate class balancing measures can help improve classifier performance in supervised and semi-supervised tests [22]. The limitations of existing machine learning-based approaches for COVID-19 detection are discussed by López-Cabrera et al. [23] Tsiknakis et al. [24] present an artificial intelligence-based framework for COVID-19 screening that can be interpreted using CXR imagery.

The research aims to make the following main contribution based on the shortcomings of several researchers' classification methods:

1) To augment the limitations of the dataset generated from several hospitals in Semarang, especially for the types of COVID-19 and Tuberculosis (TBC) diseases.

2) Perform classification Performance comparisons with new approach machine learning models without selecting features from each image, so that the best machine learning algorithm model is produced based on its Performance.

III. METHODOLOGY

A. Dataset

Fig. 1 depicts the imbalance in the number of X-ray images for each class. Normal cases outnumber COVID-19 and Tuberculosis cases. This was seen as a disadvantage by researchers because the model worked better in major classes (Normal) than in minor classes. FP (normal classified as COVID-19 or TB) would be expensive, but it was much less expensive than FN (normal classified as COVID-19 or TB), which could be fatal. Therefore, the researchers addressed the class imbalance. When building a reliable image classifier with very little training data, image augmentation was usually required to improve deep network Performance. Image augmentation generated training images artificially by using various processing methods or combinations of processing methods on each training sample, such as random rotation, shift, sliding, and flipping [15]. To accomplish this, Image Data Generator would automatically label all data inside the COVID-19 or TBC folder as COVID-19 or TBC, and all data inside the Normal folder as Normal. As a result, the data was easily ready to be fed into machine learning.

B. Data Augmentation

Data augmentation was used by researchers to increase the number of samples and diversify the images in terms of position, orientation, brightness, and so on. In this study, the augmentation technique was the best method for adding data without having to look for primary data obtained at the hospital using strict procedures. Several augmentation techniques were employed in this study, including rotation range= 40, shear range = 0.2, zoom range = 0.2, width shift range = 0.2, height shift range = 0.2, and horizontal flip. Fig. 2 shows the X-ray images for each class. Fig. 3 depicts the augmentation results.

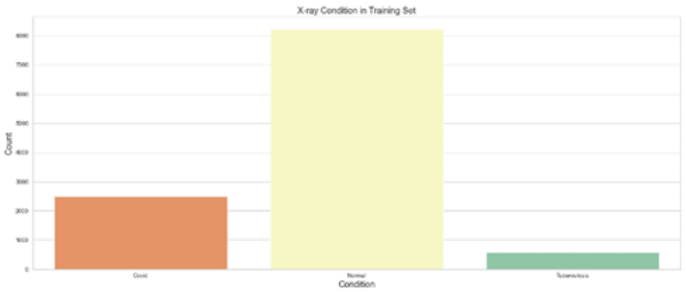


Fig. 1. Distribution of the number of dataset for lung diseases.

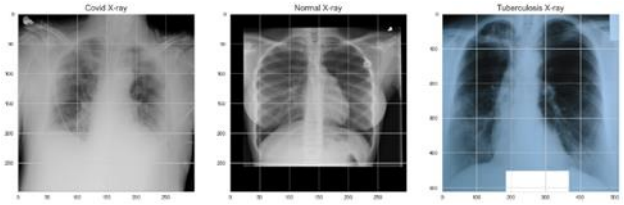


Fig. 2. X-ray images for each class (Normal, COVID-19, Tuberculosis (TB)).

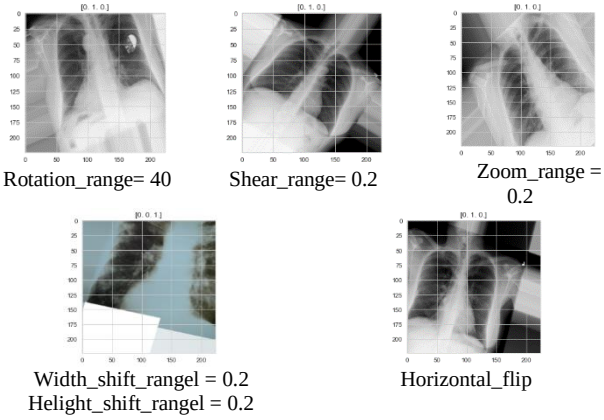
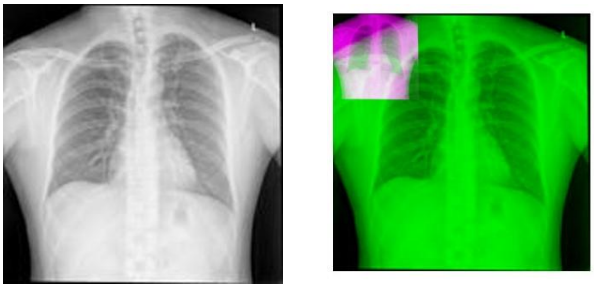


Fig. 3. X-ray image augmentation results.

As shown in Table I, we collected lung disease images from several hospitals in Semarang for three classes: Normal, COVID-19, and Tuberculosis (TB) and divided them into training and testing sets.

TABLE I. DATASET DISTRIBUTION TABLE FOR EACH CLASS

Class	Training Set	Testing Set	Total
Normal	8,454	476	8,930
COVID-19	2,363	490	2,853
Tuberculosis	474	18	492
Total	11,291	984	12,275



Original Image 224x224x3 Resize 100 x 100x3

Fig. 4. Resizing the original image into a resized image.

Fig. 4 depicts the X-ray image dataset, which is divided into three categories: normal, COVID-19, and tuberculosis (TB). The image has a resolution of 224x224 pixels and a bit depth of 16 bits [0-65535].

Fig. 4 depicts the change in dimensions of an X-ray image from 224x224x3 channels to 100x100x3 channels for computational time requirements, because the classification program must convert the images into a 2-dimensional matrix so that they can be processed in machine learning.

C. Research Model

The methodology for the study is depicted in Fig. 5. First, each class of lung diseases images was labelled. Image data was transformed into a 2D array. The next step was to reduce the size of the 2D image from 224x224x3 channels to 100x100x3 channels. The augmentation process was used to avoid unbalanced data when reproducing existing data. This study's augmentation process included a rotation range of 40, a shear range of 0.2, a zoom range of 0.2, a width shift range of 0.2, and a height shift range of 0.2. Furthermore, once the data had ben balanced, the next step was to feed 2D array data into the machine learning model to obtain classification performance so that each machine learning model with the highest accuracy, precision, recall, and f1 in each data training and testing can be evaluated.

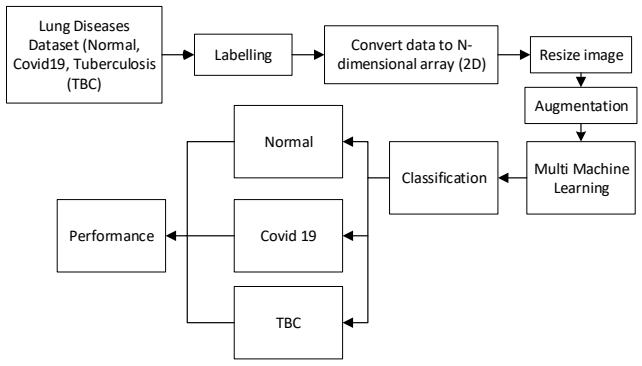


Fig. 5. Research model.

IV. Machine Learning

A. K- Nearest Neighbor

The algorithm K-Nearest Neighbors is not parametric. On the basis of a given problem or data set, learning and predictive analysis are Performed. With no dataset assumptions, the KNN classification model is a pure prediction based on neighboring data values. The letter 'K' in KNN stands for the number of nearest neighbor values in the data. The KNN algorithm classifies a given dataset based on 'K,' or the number of nearest neighbors [25]. The Euclidian distance calculation method is the most commonly used. Equation (1) gives the Equation for Euclidean calculations.

$$Euclidean_{i,j} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (1)$$

The KNN algorithm is composed of four steps. The distance from the new data to all data is calculated in the first step. The distances are then sorted in the second step. The third step finds the smallest k value, and the final step determines the class.

B. Extra Tree

Extra Trees is an ensemble machine learning approach that trains multiple decision trees and combines the results of a group of decision trees to generate predictions. However, there is a distinction between Extra Trees and Random Forest. To ensure that the decision trees are sufficiently distinct, Random Forest used bagging to select different variations of the training data. Extra Trees, on the other hand, used the entire data set to train decision trees. Therefore, it randomly selected values to split features and created child nodes in order to ensure sufficient differences between each decision tree. Extra trees can reduce model bias by using the entire dataset, which is the default setting and can be changed. On the other hand, randomizing the feature values to be split increased both bias and variance.

C. Support Vector Machine (SVM)

This stage involved using the SVM classification method to generate a classifier in the form of a feature vector, which was then used to generate predictions for testing. Following the completion of the preceding stages, the extraction results were used to generate an SVM classification model. The distance of the vector that had been mapped would be calculated. The greatest distance would be used as the vector's class separator. Then a hyperplane was added to separate the two classes. The most important requirement for developing an SVM classification model was to convert documents into vector form. Value variations were used to find values that were as accurate as possible. This procedure was necessary in order to convert the test document into a vector. The following step was to take the test document data, which was commonly referred to as a vector, and insert it into the previously created SVM model. The hyperplane Equation was a classification Equation, as shown by Equation (2), with the classification parameters *w* and *b* as the weight and bias values, as shown by Equations (3) and (4).

$$f_{svm}(x) = w.x + b \quad (2)$$

$$w = \sum_1^N a_i\, y_i x_i \quad (3)$$

$$b = -\frac{1}{2}\,(w.x^+ + w.x^-) \quad (4)$$

Where N is the amount of data and ai is the weight coefficient value for each pair of data points and labels (xi,). SVM is also good at dataset management because it finds

hyperplanes using trick kernels, one of which is the Linear Kernel shown in the Equation (5).

$$K(x_i, x_j) = x_i, x_j \quad (5)$$

From the kernel results obtained, SVM creates a classification Equation that is adjusted to the kernel used, such as Equation (6) with class classification values based on Equation (7) [25].

$$f_{svm}(x) = \sum a_i y_i K(x_i, x) + bi \in N \quad (6)$$

$$class = \begin{cases} 1, fsvm \geq 0 \\ -1, other\ than\ that \end{cases}$$

D. Random Forest

Random forest is a classification method that consists of a collection of classification trees. For example {h(x,Θk), k = 1, …} where {Θk} is a random vector that is independently distributed and each tree chooses the class that has the most number of data (majority cote). Suppose an ensemble $h_1(x)$, $h_2(x)$, …., $h_k(x)$ with training data is randomly selected from the distribution of random vectors Y and X, the margin function (mg(X,Y)) of the random forest is defined as follows:

$$mg(XY) = \frac{\sum_{k=1}^{K} I(h_k(X)=Y}{K} - max_{j \neq Y} \left[\frac{\sum_{k=1}^{k} I(h_k(X)=j)}{K}\right] \quad (7)$$

where, I is the indication function and K is the number of trees. The margin function is used to calculate the level of the number of votes in X and Y, as well as the average vote from other classes [26].

V. Result and Analysis

This study used a primary dataset with a data ratio of 70:30 for each class, as shown in Table II. A total of 11291 data points were generated in each training data set by producing the Extra Trees model with Estimator = 500 and values for precision, recall, F1-score, and accuracy of 100%, followed by other machine learning models such as SV and Random Forest. Table III shows that the Extra Trees model has the best performance value when compared to other machine learning models, with a total of 984 data points for testing.

Fig. 6(a) shows the prediction column, and thus the row must be the actual value. In the training data, the main diagonal (8030, 2135, 4070) of the confusion matrix of the Kneighbors model (n neighbors=3) gives the correct prediction. When the actual and predicted values from the model are the same, this is the case. The actual normal number is on the first line. The model predicts 8030 cases, 358 of which are COVID-19 normal and 66 of which are Tuberculosis normal. The actual COVID-19 number can be found in the second row. The model predicts COVID-19 correctly and incorrectly, with 201 COVID-19 becoming normal and 27 COVID-19 becoming TB. Tuberculosis is in the third row. The model predicts that 470 of them will correctly predict TB, 1 will become normal, and 3 will become COVID-19.

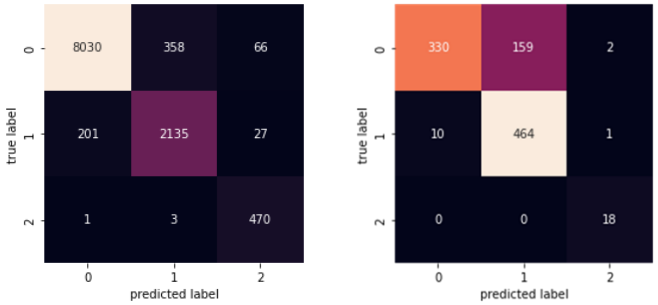
Fig. 6(b) depicts the prediction column, and thus the row must be the actual value. The main diagonal (330, 464, 18) for the confusion matrix of the Kneighbors model (n neighbors=3) testing data gives the correct prediction. This is the case when the actual and predicted values from the model are the same. The first line contains the true normal number. According to the model, 330 of them correctly and incorrectly predict 159 normal to COVID-19 and 2 normal to be Tuberculosis. The actual COVID-19 number is in the second row. The model predicts that 464 of them will correctly and incorrectly predict COVID-19, that 10 COVID-19 will become normal, and that 1 COVID-19 will become TB. Tuberculosis is in the third row. The model predicted that 18 of them would be correct and 1 would be normal.

TABLE II. Comparison of Training Data Performance Classification with a Comparison of Training and Testing Data 70:30

Training				
Classification Models	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Acc (%)**
Kneighbors (n_neighbors=3)	89	95	92	94
Kneighbors (n_neighbors=5)	84	93	88	92
Kneighbors (n_neighbors=7)	81	92	85	91
ExtraTrees (n_estimators=500)	100	100	100	100
SVM	100	100	100	100
Random Forest(n_estimators=1000)	100	100	100	100

TABLE III. Comparison of Data Testing Performance Classification with a Ratio of 70:30

Testing				
Classification Models	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Acc (%)**
Kneighbors (n_neighbors=3)	76	77	76	80
Kneighbors (n_neighbors=5)	76	79	76	80
Kneighbors (n_neighbors=7)	74	80	75	79
ExtraTrees (n_estimators=500)	89	88	87	85
SVM	83	74	77	80
Random Forest(n_estimators=1000)	86	88	85	83



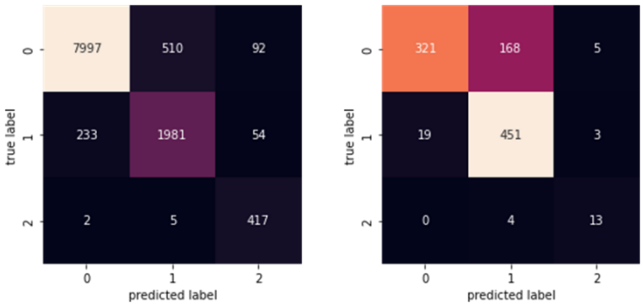
Confusion matrix training for model Kneighbors (n_neighbors=3)

(a)

Confusion matrix testing for model Kneighbors (n_neighbors=3)

(b)

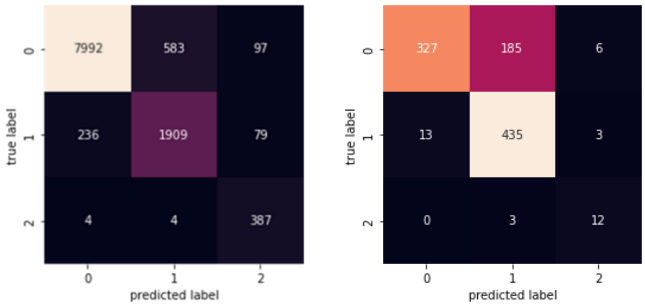
Fig. 6. Confusion matrix for the training model and testing the Kneighbors model (n_neighbors=3).



Confusion matrix training for model Kneighbors (n_neighbors=5) (a)

Confusion matrix testing for model Kneighbors (n_neighbors=5) (b)

Fig. 7. Confusion matrix for the training model and testing the Kneighbors model (n_neighbors=5).



Confusion matrix training for model Kneighbors (n_neighbors=7) (a)

Confusion matrix testing for model Kneighbors (n_neighbors=7) (b)

Fig. 8. Confusion matrix for the training model and testing the Kneighbors model (n_neighbors=7).

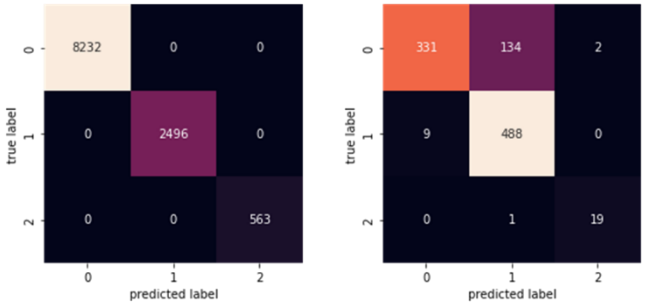
Fig. 7(a) depicts the prediction column, implying that the row must represent the actual value. The main diagonal (7997, 1981, 417) for the training data gives the correct prediction for the confusion matrix of the Kneighbors model (n neighbors=5). In this case, the actual and predicted values from the model are the same. The standard number appears on the first line. The model correctly predicts 7997 of them, 510 of which are normal to COVID-19 and 92 of which are normal to Tuberculosis. The second row contains the actual COVID-19 number. The 1981 predictive model included correctly and incorrectly predicting COVID-19, with 233 COVID-19 being normal and 52 COVID-19 being Tuberculosis. The third row is Tuberculosis. The model predicted that 417 of them would correctly predict TB, 2 TB would be normal, and 5 TB would be COVID-19.

Fig. 7(b) depicts the prediction column, and thus the row must be the actual value. For the Kneighbors model matrix confusion (n neighbors=5) data testing, the main diagonal (321, 415, 13) yields the correct prediction. When the actual and predicted values from the model are the same, this is the case. The actual normal number is on the first line. The model predicts that 321 of them will be correct, with 168 normal to COVID-19 and 5 normal to Tuberculosis. The actual COVID-19 number is in the second row. The model predicts that 451 of them will correctly predict COVID-19, 19 will become normal, and 3 will develop Tuberculosis. Tuberculosis is in the third row. The model predicts that 13 of them will correctly predict TB, and four of them will become COVID-19.

Fig. 8(a) depicts the prediction column, and thus the row must be the actual value. In the training data, the main diagonal (7992, 1909, 387) for the confusion matrix of the Kneighbors model (n neighbors=7) gives the correct prediction. When the actual and predicted values from the model are the same, this is the case. The actual normal number is on the first line. The model correctly predicts 7,992 of them, 583 of which are normal to COVID-19 and 97 of which are normal to Tuberculosis. The COVID-19 number is in the second row. The model correctly predicts COVID-19 for 1909 of them, 236 COVID-19 becomes normal, and 79 COVID-19 becomes TB. Tuberculosis is the third row. The model correctly predicted 387 of them, with 4 TB being normal and 4 TB being COVID-19.

Fig. 8(b) depicts the predicted column, and thus the row must be the actual value. For the Kneighbors model matrix confusion (n neighbors=7) data testing, the main diagonal (327, 435, 12) yields the correct prediction. When the actual and predicted values from the model are the same, this is the case. The actual normal number is on the first line. The model predicts that 327 of them will correctly predict, 185 will be normal for COVID-19, and 6 will be normal for Tuberculosis. The actual number of COVID-19 is shown in the second row. The model predicts that 435 of them will correctly predict COVID-19, 13 will become normal, and 3 will develop Tuberculosis. Tuberculosis is in the third row. The model predicts that 12 of them will correctly predict TB and 4 will correctly predict COVID-19.

Fig. 9(a) depicts the prediction column, implying that the row must represent the actual value. The training data for the ExtraTrees model confusion matrix (n estimators=500) gives the correct prediction along the main diagonal (8232, 2496, 563). In this case, the model's actual and predicted values are the same. The true normal number is found on the first line. The model correctly predicts 8232 of them. The actual COVID-19 number is in the second row. 2496 of them correctly predict COVID-19, according to the model. Tuberculosis is in the third row. The model correctly predicted TB in 563 of them.

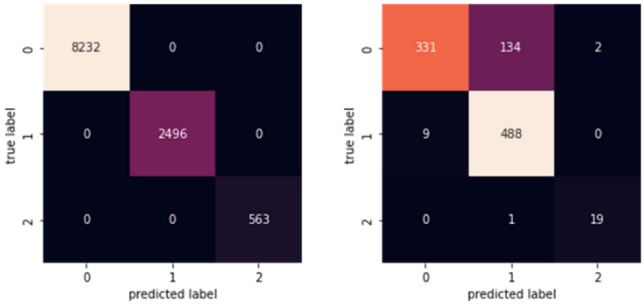


Confusion matrix training for ExtraTrees model (n_estimators=500) (a)

Confusion matrix testing for ExtraTrees model (n_estimators=500) (b)

Fig. 9. Confusion matrix for training and testing the ExtraTrees model (n_estimators=500).

Fig. 9(b) shows the predicted column, and thus the row must be the actual value. The main diagonal (331, 488, 19) for the ExtraTrees model confusion matrix (n estimators=500) testing data provides the correct prediction. This is the case when the model's actual and predicted values are the same. The actual normal number appears on the first line. The model predicts that 331 of them will correctly predict COVID-19, 134 will correctly predict Tuberculosis, and 2 will correctly predict Tuberculosis. The COVID-19 number is in the second row. The model predicts that 488 of them will correctly predict COVID-19 and 9 will correctly predict COVID-19 as normal. Tuberculosis is the third row. 19 of them correctly predict TB, with 1 TB being COVID-19, according to the model.



Confusion matrix training for Support Vector Machine (SVM) model
(a)

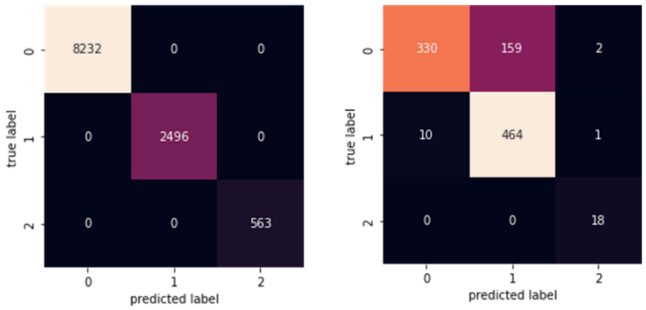
Confusion Matrix testing for Support Vector Machine (SVM) model
(b)

Fig. 10. Confusion matrix for training and testing models for Support Vector Machine (SVM).

Fig. 10(a) shows the prediction column, and thus the row must be the actual value. The Support Vector Machine (SVM) data model confusion matrix's main diagonal (8232, 2496, 563) yields correct predictions. This is the case when the model's actual and predicted values are the same. The true normal number is found on the first line. The model correctly predicts 8232 of them. The actual COVID-19 number is in the second row. 2496 of them correctly predict COVID-19, according to the model. Tuberculosis is in the third row. The model correctly predicted TB in 563 of them.

Fig. 10(b) shows the predicted column, implying that the row must be the actual value. For the Confusion Support Vector Machine (SVM) model testing data, the main diagonals (331, 488, 19) provide the correct predictions. When the model's actual and predicted values are the same, this is the case. The true normal number is on the first line. The model predicts that 331 of them will correctly predict, 134 will become COVID-19, and 2 will become Tuberculosis. The actual COVID-19 number is in the second row. The model predicts that 488 of them correctly predict COVID-19, with 9 of them becoming normal. Tuberculosis is in the third row. The model predicts that 19 of them will correctly predict TB, with one TB being COVID-19.

Fig. 11(a) shows the predicted column, and thus the row must be the actual value. The training data for the Random Forest model confusion matrix (n estimators=500) gives the correct prediction along the main diagonal (8232, 2496, 563). This is the case when the model's actual and predicted values are the same. The true normal number is found on the first line. The model correctly predicts 8232 of them. The actual COVID-19 number is in the second row. 2496 of them correctly predict COVID-19, according to the model. Tuberculosis is in the third row. The model correctly predicted TB in 563 of them.



Confusion matrix training for random forest model (n_estimators=500)
(a)

Confusion matrix testing for random forest model (n_estimators=500)
(b)

Fig. 11. Confusion matrix for training and testing models for random forest (SVM).

Fig. 11(b) depicts the prediction column, and thus the row must be the actual value. The main diagonals (330, 464, 18) for the Random Forest Confusion model (n estimators=500) test data give the correct predictions. This is the case when the model's actual and predicted values are the same. The actual normal number appears on the first line. The model predicts that 330 of them will be correct, with 159 being normal and two being Tuberculosis. The COVID-19 number is in the second row. The model predicts that 464 of them will correctly predict COVID-19, 10 will be normal, and 1 will be Tuberculosis. Tuberculosis is the third row. The model correctly predicted TB for 18 of them.

Table IV shows comparison of the accuracy values with several researchers. Bakir et al applied deep learning techniques to detect pneumonia from X-ray images. They used an artificial neural network (ANN) model to classify bacterial, viral, and healthy lungs into multiple classes. His proposed ANN model with ResNet feature extraction, in multi-class classification he achieved 81.67% classification accuracy [27]. Kim, Sungyeup et al used chest X-ray (CXR) images to diagnose three classes, normal, pneumonia and peumothorax. The method used uses a deep learning model (EfficientNet V2-M) to produce an accuracy value of 82.15% [28].

TABLE IV. COMPARISON OF ACCURACY VALUE WITH OTHER RESEARCHERS

Prior Work	Model	Acc (%)
Bakir et al [27]	ANN model	81.67%
Kim, Sungyeup et al [28]	EfficientNet V2-M	82.15%.
Proposed Method	ExtraTrees	85%

VI. CONCLUSION

Through the use of multiple machine learning techniques, this study contributes to the multi-class classification of lung diseases. The research methodology used in this study is an experimental multi-class classification of lung disease using an augmentation process to obtain balanced data.

The experimental results produce the Extra Trees classification which has Precision, Recall, F-Score, and Accuracy score of 100% for training, and testing data, 89% for Precision, 88% for Recall, 87% for Fi-Score, and 85% for Accuracy higher than the Performance of other machine learning models such as Kneighbors, Support Vector Machine (SVM), Random Forest and more effective in Classification of lung diseases. Comparison with other researchers shows that the proposed model has a higher accuracy value compared to other models.

In further research, the dataset as a whole is still not large enough and of low quality for highly accurate and useful deep learning results that can be used as benchmarks for the identification of lung disease types by viewing X-rays. More high-quality images are needed to increase the accuracy of average ratings to higher levels in multiple-class classification. The data is expected to grow over time, enabling better classification results.

ACKNOWLEDGMENT

This research supported by Institutional internal competitive grants 2022 from Universitas Semarang No: 008/USM.H7.LPPM/L/2022

REFERENCES

[1] Saputra, "SISTEM PAKAR IDENTIFIKASI PENYAKIT PARU-PARU PADA MANUSIAMENGGUNAKAN PEMROGRAMAN VISUAL BASIC 6.0," J. Teknol. DAN Inform., vol. 1, no. 3, pp. 202–222, 2011.

[2] A. Mardhiyah; AgusHarjoko, "Metode Segmentasi Paru-paru dan Jantung Pada Citra X- Ray Thorax," IJEIS (Indonesian J. Electron. Instrum. Syst., vol. 1, no. 2, pp. 35–44, 2012.

[3] B. Chen, J. Li, X. Guo, and G. Lu, "Biomedical Signal Processing and Control DualCheXNet : dual asymmetric feature learning for thoracic disease classification in chest X-rays," Biomed. Signal Process. Control, vol. 53, p. 101554, 2019, doi: 10.1016/j.bspc.2019.04.031.

[4] C. M. Sharma, L. Goyal, V. M. Chariar, and N. Sharma, "Lung Disease Classification in CXR Images Using Hybrid Inception-ResNet-v2 Model and Edge Computing," J. Healthc. Eng., vol. 2022, pp. 1–15, 2022.

[5] A. Abbas and M. M. Abdelsamea, "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network," Appl. Intell., pp. 854–864, 2021.

[6] A. Souid, N. Sakli, and H. Sakli, "applied sciences Classification and Predictions of Lung Diseases from Chest X-rays Using MobileNet V2," Appl. Sci., vol. 11, no. 2751, pp. 1–16, 2021.

[7] M. Zak, "Classification of Lung Diseases Using Deep Learning Models," Concordia University, Montr´eal, Qu´ebec, Canada, 2019.

[8] A. Bhandary et al., "Deep-Learning Framework to Detect Lung Abnormality – A study with Chest X-Ray and Lung CT Scan Images," Journla Pre-proof, 2019, doi: 10.1016/j.patrec.2019.11.013.

[9] M. Z. P. Emtiaz Hussain , Mahmudul Hasan , Md Anisur Rahman , Ickjai Lee , Tasmi Tamanna, "CoroDet : A deep learning based classification for COVID-19 detection using chest X-ray images," Chaos, Solitons and Fractals, vol. 142, p. 110495, 2021, doi: 10.1016/j.chaos.2020.110495.

[10] S. Albahli, "A Deep Neural Network to Distinguish COVID-19 from other Chest Diseases Using X-ray Images," Curr. Med. imaging, vol. 17, pp. 109–119, 2021, doi: doi:10.2174/1573405616666200604163954.

[11] J. C. Souza et al., "Computer Methods and Programs in Biomedicine An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks," Comput. Methods Programs Biomed., vol. 177, pp. 285–296, 2019, doi: 10.1016/j.cmpb.2019.06.005.

[12] R. M. Pereira, D. Bertolini, L. O. Teixeira, C. N. Silla, and Y. M. G. Costa, "Computer Methods and Programs in Biomedicine COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios," Comput. Methods Programs Biomed., vol. 194, p. 105532, 2020, doi: 10.1016/j.cmpb.2020.105532.

[13] B. Ergen and Z. Cömert, "A Deep Feature Learning Model for Pneumonia Detection Applying a Combination of mRMR Feature Selection and Machine Learning Models," IRBM, vol. 1, pp. 1–11, 2019, doi: 10.1016/j.irbm.2019.10.006.

[14] P. Ranjan and R. Janardhanan, "Pneumonia Identification in Chest X-Ray Images Using EMD," in Trends in Communication , Cloud , and Big Data, no. January, 2020, pp. 87–98.

[15] A. Géron, Hands-On Machine Learning with Scikit-Learn. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472., 2017.

[16] G. Liang, C. Greenwell, and S. Member, "Contrastive Cross-Modal Pre-Training : A General Strategy for Small Sample Medical Imaging," IEEE J. Biomed. Heal. Informatics, vol. XX, no. X, pp. 1–10, 2021.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet : Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9049–9058.

[18] Y. Tang, Y. Tang, Y. Zhu, and J. Xiao, "A Disentangled Generative Model for Disease Decomposition in Chest X-rays via Normal Image Synthesis," Med. Image Anal., p. 101839, 2020, doi: 10.1016/j.media.2020.101839.

[19] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-mahmood, "SEMI-SUPERVISED LEARNING WITH GENERATIVE ADVERSARIAL NETWORKS FOR CHEST X-RAY CLASSIFICATION WITH ABILITY OF DATA DOMAIN ADAPTATION," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, no. Isbi, pp. 1038–1042.

[20] S. Albahli, H. T. Rauf, A. Algosaibi, and V. E. Balas, "AI-driven deep CNN approach for multi- label pathology classification using chest," PeerJ Comput. Sci., pp. 1–17, 2021, doi: 10.7717/peerj-cs.495.

[21] I. M. Baltruschat and H. Nickisch, "Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification," Sci. Rep., pp. 1–11, 2019, doi: 10.1038/s41598-019-42294-8.

[22] S. Calderon-ramirez et al., "Correcting data imbalance for semi-supervised COVID-19 detection using X-ray chest images," Appl. Soft Comput. J., no. January, 2020.

[23] J. Daniel et al., "Current limitations to identify COVID - 19 using artificial intelligence with chest X - ray imaging," Health Technol. (Berl)., pp. 411–424, 2021, doi: 10.1007/s12553-021-00520-2.

[24] A. H. K. and K. M. Nikos Tsiknakis, Eleftherios Trivizakis, Evangelia E. Vassalou, Georgios. Papadakis, Demetrios A. Spandidos5, Aristidis Tsatsakis, Jose Sánchez-García, Rafael López-González, Nikolaos Papanikolaou, "Interpretable artificial intelligence framework for COVID-19 screening on chest X-rays," Exp. Ther. Med., vol. 20, pp. 727–735, 2020.

[25] M. E. A. R. Risha Ambar Wati , Hafiz Irsyad, "Klasifikasi Pneumonia Menggunakan Metode Support Vector Machine," J. Algoritm., vol. 1, no. 1, pp. 21–32, 2020.

[26] D. Of, P. From, X. I. Using, and D. Learning, "DETECTION OF PNEUMONIA FROM X-RAY IMAGES USING DEEP LEARNING," J. Sci., vol. 52, pp. 419–440, 2023.

[27] Halit BAKIR; Semih OKTAY; Timuçin Emre TABARU, "DETECTION OF PNEUMONIA FROM X-RAY IMAGES USING DEEP LEARNING," J. Sci., vol. 52, pp. 419–440, 2023.

[28] S. Kim, B. Rim, S. Choi, A. Lee, and S. Min, "Deep Learning in Multi-Class Lung Diseases' Classification on Chest X-ray Images," Diagnostics, vol. 12, no. 915, pp. 1–24, 2022.