Comparison of machine learning algorithms for chest X-ray image COVID-19 classification

Comparison of machine learning algorithms for chest X-ray image COVID-19 classification

Samsir Samsir[1*], Jimmi Hendrik P. Sitorus[2], Zulkifli[3], Zuriani Ritonga[4], Fitri Aini Nasution[5], Ronal Watrianthos[6]

[1,6] Informatics Department, Universitas Al Washliyah Labuhanbatu, Indonesia
[2] AMIK Parbina Nusantara, Pematang Siantar, Indonesia
[3] STIA Setih Setio Muara Bungo, Indonesia
[4] Management Department, Universitas Labuhanbatu, Indonesia
[5] Informatic System, Institut Teknologi dan Sains Paluta, Indonesia

*samsirst111@gmail.com

Abstract. Artificial Intelligence and Machine Learning algorithms were used to identify the coronavirus (COVID-19) from X-ray photos of the chest. The authors propose a model for early coronavirus detection based on image filtering strategies and a hybrid feature selection model in this analysis. Traditional statistical and machine learning methods are used to derive these attributes from CT images. The Confusion Matrix for infected COVID-19 patients and regular patients was obtained using Support Vector Machine and K-Nearest Neighbor to classify the features chosen. The output of the two approaches can be compared. The various techniques' performance shows that the Support Vector Machine achieves the highest precision of 97% compared to the K-Nearest Neighbor with a precision of 86%.

1. Introduction

Coronavirus (COVID-19) spreads from an infected person by saliva droplets, coughs, and sneezes. The majority of individuals infected with coronavirus (COVID-19) have minor respiratory diseases and are more prone to experience serious diseases such as cardiovascular disease, asthma, chronic respiratory diseases, and cancer. Many individuals above the age of 60 and underlying clinicians are at high risk for COVID 19[1]. The coronavirus is the fastest transmitted virus among humans as a consequence of serious acute respiratory syndrome. From the CT x-ray and the signs, a method to correctly classify the coronavirus. Two hundred fifty-three samples of infected COVID-19 patients were collected using a different source. A licensed clinical laboratory tested the blood of 49 patients and 24 patients infected. The cross-validation approach correctly detects contaminated patients with a sensitivity of 96.95 percent and a precision of 95 percent[2]. Detecting this condition from X-ray scans is also one of the fastest methods to detect patients. In early studies, infected patients have abnormalities in the chest X-rays

Artificial Intelligence and Machine Learning algorithms can provide identification of coronavirus from chest X-ray photos. Classify the files using CNN with the SoftMax classifier, SVM, and the random forest. CNN is seen in two scenarios: picture classification and graphical attribute extraction with a hybrid method. Train and measure parameters using to derived function. According to the proposed algorithm, CNN precision is 95.2 percent, which is higher than other approaches[3]. The usage of radiographic and radiology scans to detect the infection is one of the quickest strategies to diagnosis patients. Early findings indicate that the chest x-rays of COVID-19 patients are visibly

abnormal. This research uses Deep Learning models to detect COVID-19 patients using 5000 Chest X-rays from publicly accessible datasets. The technique used produces a heat map of the pulmonary area potentially infected with COVID-19. It shows that the resulting heat map includes most of the infected areas identified by accredited radiologists[4].

The authors propose a model for early coronavirus identification focused on image filtering techniques, and a hybrid feature selection model in this study extracted these attributes using statistical, and machine learning approaches[5]. Comparison of the SVM and KNN methods using the selected characteristics is the aim of this study.

2. Methodology

There are 5,863 X-Ray photos (JPEG) and two sections (Pneumonia/Normal) in the dataset repository. From the open-source Kaggle website, 80% of the dataset is used for preparation, and 20% is used for testing[6]. As part of routine, outpatient care for patients using Chest X-ray scan. Both chest x-rays have been checked for quality control first, with scans excluded from bad or unreadable. Two specialists then graded the diagnoses of the pictures until the AI scheme was accepted. A third specialist reviewed the evaluation package to make sure there were no grading mistakes. The grey levels, patch scale, measurements, and features of the X-ray images were all new[7].

Figure 1. Normal and abnormal sample [6]

2.1 Pre-processing

By adding a median filter, average filter, and histogram equalization, the preprocessing phase improves its generalization for eliminating noise and improving the contrast enhancement in the entire picture. The median filter sorts pixels in the picture and replaces them with the pixels' median in the surrounding area. The average filter smoothest the picture by reducing adjacent pixel amplitude variations and replacing a neighboring pixel's average value, including itself[8]. Furthermore, Histogram equalization increases picture contrast by stretching out the intensity spectrum, resulting in a higher resolution image with no detail loss.

2.2 Feature extraction

Different features have been widely utilized in extraction and selection processes. The HOG mechanism synthesizes dimensional distribution in the picture areas and is particularly helpful in defining deformable structures. The technique is convincing enough to calculate the histogram quickly[9].

2.3 Machine learning

Machine learning is a multidisciplinary discipline with a diverse set of science domains supporting it. Computational Statistics, whose fundamental goal is to make forecasts using machines, is closely linked to ML models' simulation[10]. It has also linked to Mathematical Optimization, a branch of statistics that deals with templates, implementations, and frameworks. Machine learning can be used to build and program explicit high-performance algorithms in a number of computing fields [11][12][13].

2.4 Support Vector Machine (SVM)

This algorithm used to characterize a single entity based on derived attributes. Any features will be extracted, and these features must be transferred through the SVM module to detect the correct entity. The SVM algorithm uses a hyperplane to segregate or add a sample to its class. The number of

features and the corresponding characteristics must be defined to optimize each function's utility and make the detection process more effective. Supporting Vector Machine provides more reliable performance, but limiting this method is that the time required for classification is more compared to other weaker classifiers. SVM library is used to facilitate module creation[14].

2.5 K-Nearest Neighbor (KNN)

Algorithm K-Nearest Neighbor is an algorithm classifying objects that are nearest to the object. The K-Nearest Neighbor algorithm categorizes new data which its class still does not know by choosing k data which are closest to the new ones. As the expected class for new data, the closest class frequency of k is selected. In general, the value of k uses the odd number such that the classification method does not have the same distance. The distance or nearness of neighbours is determined by Euclidean distance [15][16][5].

2.6 Performance Evaluation

In machine learning, the confusion matrix is more widely used to assess a classification model's efficiency. In a classification problem, the correct and incorrect results are tallied, and the performance is compared to the reference data. Accuracy, Precision, Recall, Specificity, and F1-score are some of the most popular matrices. Four statistical indices were determined to overcome the uncertainty matrix: true positive (TP), true negative (TN), false positive (FP), and false negative (FN), as shown in equation (2)(5).

$$Accuracy = \frac{(TN+TP)}{TN+TP+FN+FP} \quad (1)$$

$$Precision = \frac{(TP)}{TP+FP} \quad (2)$$

$$Recall = \frac{(TP)}{TP+FN} \quad (3)$$

$$F1 - Score = 2 * \frac{(Precision*Recall)}{(Precison+Recall)} \quad (4)$$

True Positive (TP) means they've been infected with COVID-19; True Negative (TN) means they haven't; False Positive (FP) means they've been infected with COVID-19; False Negative (FN) means they haven't been infected with COVID-19[17].

3. Result and Discussion

A chest X-ray dataset was used to predict coronavirus (COVID-19) infected patients and regular patients in this research. Machine learning algorithms such as Support Vector Machine (SVM) and K-Nearest Neighbor were used to train and validate the dataset on chest X-Ray pictures (KNN). The SVM algorithm obtained a 98 per cent accuracy, 97 per cent precision, 94 per cent recall, 94 per cent specificity, and 98 per cent F1.

Table 1. Confusion Matrix

Classifications	TP	FP	TN	FN
SVM	269	15	145	7
KNN	250	24	152	12

Table 1 reveals that, as opposed to the K-Nearest Neighbor (KNN) algorithm, the SVM produced the better predictive performance utilizing different metric indicators.

Table 2. Performance Result (%)

Classifications	Accuracy	Precision	Recall	Specificity	F1
Support Vector Machine (SVM)	98	97	100	94	98
k-Nearest Neighbors (KNN)	94	94	98	89	96

The Confusion matrix for COVID-19 infected patients and regular patients was obtained using the proposed machine learning techniques, as seen in Table 1 and Table 2. When the uncertainty matrix was analyzed, the Support vector machine identified COVID-19 contaminated patients (269 pictures) as true positive and regular images (139 images) as true negative, achieving a 98 percent accuracy score. The KNN has a 94 percent precision, with true positive images of COVID-19 of 250 and true negative images of 152.

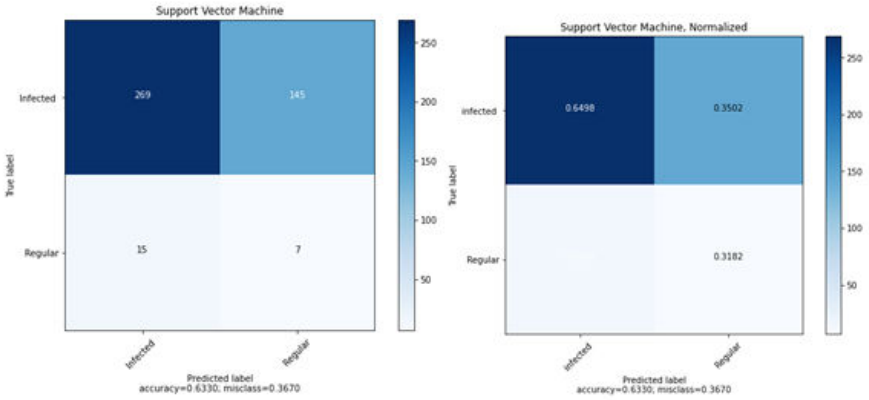

Figure 1. The confusion matrix for SVM and after normalized

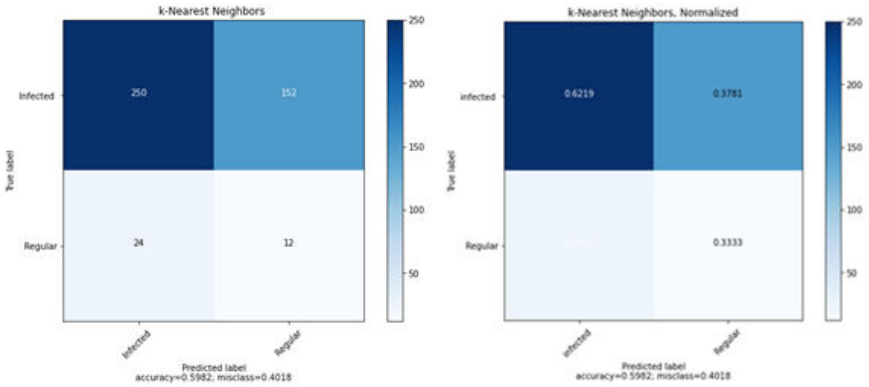

Figure 2. The confusion matrix for KNN and after normalized

Classifiers are seen by drawing a confusion matrix in Figures 1 and 2. The SVM and KNN settings are used to construct a regular, normalized uncertainty matrix. The support vector machine has the best performance model of the machine learning process, with 97 percent precision, 100 percent recall, 94 percent specificity, and 99 percent F1 score, compared to k-Nearest Neighbors, which has 86 percent precision, 87 percent recall, 73 percent specificity, and an F1-score of 86 percent, as seen in Figure 3.

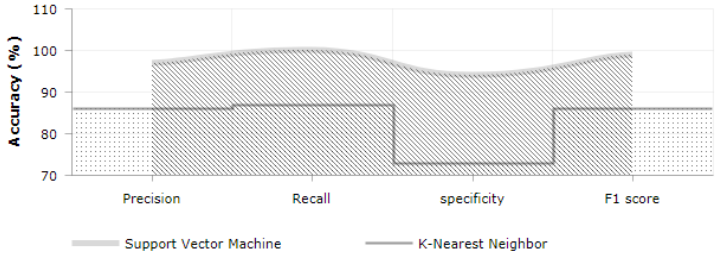

Figure 3. Comparison Performance of SVM and KNN

4. Conclusion

Among other things, early COVID-19 predictions might have ended the epidemic. This research used some devices with chest x-rays to discern between infected COVID-19 patients and standard chest X-ray pictures. The reliability of the various techniques reveals that the SVM reaches the highest precision of 97% compared to the KNN with a precision of 86%. Comparisons with other classification methods, such as Random Forest or Naïve Bayes, may need to be made in the future to use extraction techniques to increase the performance quality and aid decision-making in clinical practice.

References

[1] J. Morrison, "UM School of Medicine is First in U.S. to Test Unique RNA Vaccine Candidate for COVID-19," *University of Maryland School of Medicine*, 2020. https://www.medschool.umaryland.edu/news/2020/UM-School-of-Medicine-is-First-in-US-to-Test-Unique-RNA-Vaccine-Candidate-for-COVID-19.html (accessed Mar. 05, 2021).

[2] J. Wu *et al.*, "Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results," *medRxiv*, 2020, doi: 10.1101/2020.04.02.20051136.

[3] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Comput. Biol. Med.*, 2020, doi: 10.1016/j.compbiomed.2020.103792.

[4] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. Jamalipour Soufi, "Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning," *Med. Image Anal.*, 2020, doi: 10.1016/j.media.2020.101794.

[5] M. imad, N. Khan, F. Ullah, M. A. Hassan, A. Hussain, and Faiza, "COVID-19 Classification based on Chest X-Ray Images Using Machine Learning Techniques," *J. Comput. Sci. Technol. Stud.*, vol. 2, no. 2, pp. 01–11, 2020, [Online]. Available: https://al-kindipublisher.com/index.php/jcsts/article/view/531.

[6] P. Mooney, "Chest X-Ray Images (Pneumonia)," *kaggle.com*, 2018. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (accessed Mar. 05, 2021).

[7] D. S. Kermany *et al.*, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.

[8] A. C. Frery, "Image filtering," in *Digital Document Analysis and Processing*, 2013.

[9] C. Shu, X. Ding, and C. Fang, "Histogram of the oriented gradient for face recognition," *Tsinghua Sci. Technol.*, 2011, doi: 10.1016/S1007-0214(11)70032-3.

[10] Y. Triyanto, R. Watrianthos, Y. Sepriani, and K. Rizal, "Palm Oil Prediction Production Using Extreme Learning Machine," *Int. J. Sci. Technol. Res.*, vol. 8, no. 08, pp. 1070–1072, 2019.

[11] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," *J. Phys. Conf. Ser.*, vol. 1142, p. 012012, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012012.

[12] M. Nasution, D. Irmayani, R. Watrianthos, S. Suryadi, and I. R. Munthe, "Comparative Analysis Of Data Mining Using The Rought Set Method With K-Means Method," *Int. J. Sci. Technol. Res.*, vol. 8, no. 05, pp. 38–40, 2019, [Online]. Available: http://www.ijstr.org/final-print/may2019/Comparative-Analysis-Of-Data-Mining-Using-The-Rought-Set-Method-With-K-means-Method.pdf.

[13] R. A. Purba, S. Samsir, M. Siddik, S. Sondang, and M. F. Nasir, "The optimalization of backpropagation neural networks to simplify decision making," 2020, doi: 10.1088/1757-899X/830/2/022091.

[14] G. J. Frederick, G. Sakthivel, and Jagadeeshwaran, "Impact of Adaboost and Support Vector Machine Classifier in Automotive Sector," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 624, p. 012031, Oct. 2019, doi: 10.1088/1757-899X/624/1/012031.

[15] A. G. Pertiwi, N. Bachtiar, R. Kusumaningrum, I. Waspada, and A. Wibowo, "Comparison of performance of k-nearest neighbor algorithm using smote and k-nearest neighbor algorithm

without smote in diagnosis of diabetes disease in balanced data," *J. Phys. Conf. Ser.*, vol. 1524, p. 012048, Apr. 2020, doi: 10.1088/1742-6596/1524/1/012048.

[16] I. Kerenidis and A. Prakash, "Quantum recommendation system," 2017, doi: 10.4230/LIPIcs.ITCS.2017.49.

[17] M. Imad, N. Khan, F. Ullah, M. A. Hassan, A. Hussain, and Faiza, "COVID-19 Classification based on Chest X-Ray Images Using Machine Learning Techniques," *J. Comput. Sci. Technol. Stud.*, vol. 2, no. 2, 2020.