Majority Voting Algorithm for TB Detection: Machine Learning Approach

Nkgaphe Tsebesebe[1,2], Kelvin Mpofu[1], Sphumelele Ndlovu[1], Sudesh Sivarasu[2] and Patience Mthunzi-Kufa[1,2,3]

[1]Council for Scientific and Industrial Research, National Laser Centre, P.O Box 395, Building 46A, Pretoria 0001, South Africa.
[2]Department of Human Biology, Division of Biomedical Engineering, University of Cape Town, Cape Town 7935, South Africa
[3]School of Chemistry and Physics, University of KwaZulu-Natal, Durban 4001, South Africa

E-mail: ntsebesebe@csir.co.za

Abstract. An accurate examination of chest radiographs remains a challenge for physicians. As part of the solution, most researchers proposed the use of algorithms to perform automatic classification of chest radiographs. The current work continues with the effort to develop machine learning algorithms to perform the classification of chest radiograph images as healthy and TB infected. In this work, a combination of machine learning algorithms (i.e., a artificial neural network, a support vector machine, and decision tree algorithms) use a majority voting technique to classify chest X-ray images into healthy and TB-infected. The performance of individual algorithms is optimized by hyperparameter techniques such as grid search. In this work we develop a voting system where the three machine learning algorithms collectively classify each image and the class with the majority number of votes becomes the new label. The voting system of this work achieves an accuracy of 99% in classifying 800 chest radiograph images. The system classifies TB-infected images by 98% and noninfected images by 100%, providing high precision and recall. The voting system outperforms the individual algorithms' performances. The voting algorithm proposed in this work can be beneficial in medical facilities in providing second opinions to medical professionals by reducing the difficult task of classifying abnormalities in chest radiographs.

1. Introduction

Chronic obstructive pulmonary diseases (COPDs) are considered the third most common cause of death worldwide [1]. These are part of lung diseases that lead to respiratory problems. Respiratory or chest diseases are common diseases, including among others tuberculosis (TB), which is the primary cause of mortality around the world [1, 2]. In 2021, TB caused by Mycobacterium tuberculosis (Mtb) with a single bacterial agent claimed approximately 1.6 million lives worldwide [3, 4]. On the other hand, approximately 10 million people were reported to have developed active TB in 2020 [4]. The death toll for 2021 was an increase from the 1.4 million TB-related deaths worldwide reported in 2019 [4, 5]. As a result, TB claims more than a million lives a year, making it the leading cause of death [3, 4]. To ensure proper treatment, control, and complete recovery, early detection of TB becomes a crucial step [1]. Although there are numerous methods used to detect TB up to date, chest radiography is one of the most widely

used examination methods to perform TB diagnosis due to its relatively low cost and ease of operation.

The chest radiograph examines the thorax (chest) and identifies acute and chronic cardiopulmonary conditions to allow medical professionals to make the diagnosis [4]. In X-ray diagnosis, the partly transmitted and absorbed radiation is used to produce an X-ray image [6]. The images show detailed vibrations in transmitting the radiation as a result of structural objects of varying thickness, density, and atomic composition [6]. These chest radiographs contain detailed information on the health of patients, hence accurate examination of the information is important to produce medical reports of patients [4, 6]. Conventional methods of diagnosis involve physicians analyzing the images and making conclusions. However, the accurate examination of the information in detail remains a challenge to physicians [4]. Arshia Rehman et. al. [4] reported that it is a very difficult task for clinical radiologists to classify abnormalities on chest radiographs. As part of the solution, many researchers proposed the use of algorithms to perform automatic classification of chest radiograph images [4]. The approach offers multiple advantages, such as reduced diagnosis errors, localization of skeptical regions, and overcoming human bias [4].

The current work continues with the effort to develop machine learning algorithms to perform the classification of chest radiograph images as healthy and infected with TB. The approach presented in this paper makes use of the majority voting technique where three machine learning models (support vector machine, decision tree, and artificial neural network) are combined and a prediction of each model on the same x-ray image is considered a vote. The majority vote count was considered to be a final prediction on a particular image under examination. The aim of this technique was to improve the overall classification performance of the algorithms. Collective voting has shown to outperform individual algorithms.

2. Methodology

2.1. Machine Intelligence Library

The traditional models, i.e., the support vector machine and the decision tree of the current study were implemented from scikit-learn [7]; the deep learning model, i.e., the artificial neural network was implemented from Keras [7] with Google TensorFlow [8]. All the models operated with the aid of some scientific computing libraries, including numpy [9] and matplotlib [10] for mathematical operations and data visualization.

2.2. Dataset and data pre-processing

This study used publicly available chest radiograph images from the Kaggle TB database [11]. In figure 1 are samples of chest X-ray images from the database. The images are labeled; hence the current study follows the supervised machine learning approach. That is, the labeled data sets were used to train the algorithms to make accurate predictions. The work starts with 1206 chest radiographs, 800 infected with TB and 400 healthy. All the images are resized into 200 x 200 pixels, then flattened into a 1-dimensional array with 40000 shapes, from which the array contained information about each image in a grid of values representing image features. The features were then scaled to the maximum feature of 1 and a minimum of 0 in the array. Principal component analysis was utilized to perform feature selection considering a 98% variance ratio to detect sensitive information from the images while minimizing information loss. The data was split so that 80% of the chest X-ray images were for training the machine learning models and 20% of the images were for testing the model. Lastly, machine learning models were trained on the data set.

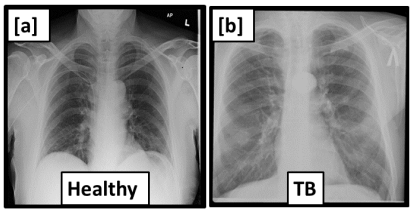

Figure 1. The samples of chest X-ray images used in this work.

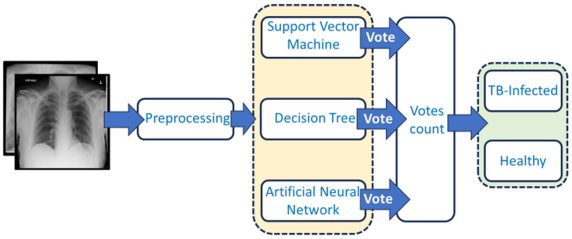

Figure 2. A graphic representation of the voting algorithm.

2.3. Majority Voting Algorithms

This study considers three machine learning models to cast predictions as votes, where the majority votes determine the prediction of each X-ray image. A representation of the voting algorithm is displayed in Figure 2.

2.4. Support vector machine

Support vector machines belong to a family of generalized linear classification [12]. They are a set of supervised learning methods that can solve pattern classification and regression problems [12]. The SVM work by minimizing the classification and maximizing the geometric margin. This happens by mapping the input vector to a higher-dimensional space from which the maximal separating hyperplane is generated [12]. Conversely, parallel hyperplanes are generated on the side of the hyperplane that separates the data. In the current work, the kernel function which transforms the data into a higher-dimensional space to perform separation of the data was a linear function. The hyperplane took the form of:

$$m \cdot x + c = 0 \tag{1}$$

where c is a scalar and m is a number of images, and x are features. Parallel hyperplanes are defined as:

$$m \cdot x + c = 1 \tag{2}$$

$$m \cdot x + c = -1 \tag{3}$$

Since the data in this work are linearly separable, the hyperplanes were selected so that there were no points between them. The distance between the hyperplanes was determined as:

$$\frac{2}{|m|} \tag{4}$$

The support vector machine in this case minimized $|m|$ from

$$m \cdot x_i - c \geq 1 \; or \; m \cdot x_i - c \leq 1, \; 1 \leq i \leq n \tag{5}$$

2.5. Decision tree

A decision tree is a type of supervised learning model that adopts a flow chart tree structure, where every internal node represents a test on an attribute, the branches denote the prediction of the test and the leaf nodes represent labels of the classes [12, 13]. This work implemented a decision tree algorithm with an information gain function to facilitate data splitting. The function considered class probability and log_2 of class probability as follows:

$$information \; gain \; or \; Entropy = \sum_{n=2}^{C} -p_i \times log_2(p_i) \tag{6}$$

Where p represents the probability of samples in a class, C represents a number of classes and i is a number of samples. The decision tree algorithm of this work used 8 minimum number of chest X-ray images to perform the splitting of the internal note to a maximum tree depth of 17.

2.6. Artificial neural network

This study implemented a feedforward neural network with two different activation functions: the rectified linear unit (ReLU) and the sigmoid in the hidden layer, and the classification layer, respectively. The activation function ReLU possesses biological and mathematical information that helps improve deep neural network training [14]. The functionality of the ReLU activation function depends on the thresholding values at 0. It computes the values from the following function:

$$f(x) = max(0, x) \tag{7}$$

The function output 0 when $x < 0$ implying the neuron is not activated in the case and output linear function when $x \geq 0$, neurons are activated in this case. Sigmoid activation on the classification layer was defined as:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

Where e is the exponential function and x are the features. The function translates the inputs range from $(-\infty; +\infty)$ to a range in $[0; 1]$ [15]. This enabled a prediction of either TB-infected or healthy. The artificial neural network of this work was implemented sequentially with densely connected layers. The model had only one hidden layer consisting of 64 neurons and one neuron at the classification layer to facilitate a binary classification. The output of the hidden layer was normalized from a normalization batch. The model was optimized by Adam optimizer with a learning rate of 0.18755133602737314 and a loss function of categorical crossentropy. The model was trained with 100 chest X-ray images at the time, for 80 training cycles.

3. Results and Discussion

3.1. Performance matrices

The performance of the majority-voting algorithm depends mainly on the performance of each algorithm casting votes in the process. Models in this current paper are evaluated for over-fitting and under-fitting before they can be implemented in the majority voting. Table 1 shows the performance scores of the models during testing. The support vector machine and artificial neural network both achieve 97.93% accuracy during testing, while the decision tree falls behind by 4.13% to 93.8%. The models were grouped to cast votes, where the majority votes on each individual chest radiograph image was considered a prediction of the voting algorithm. The performance of the voting algorithm was evaluated from the ability of the algorithm to classify 800 images as TB-infected or healthy. The algorithm correctly classified 596 images as TB-infected and 200 images as healthy; it incorrectly classified 4 images as healthy while they were TB-infected. This performance is shown from the confusion matrices in Figure 4.

Table 1. Performance scores of the models during testing.

Algorithms	Testing Score (%)
`Support Vector Machine`	97.93
`Decision Tree`	93.80
`Artificial Neural Network`	97.93

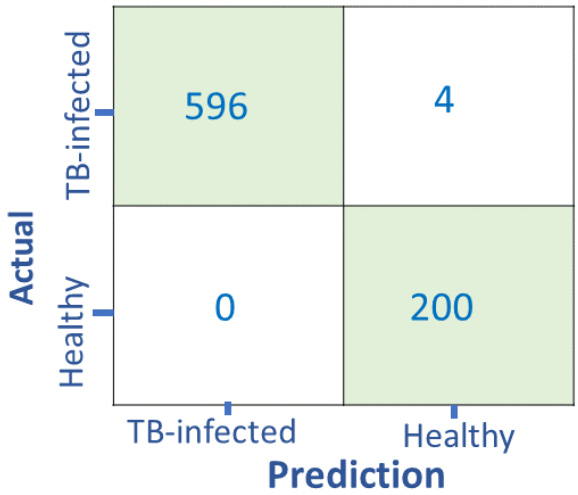

Figure 3. A performance metrics of the voting algorithm

3.2. Clinical indicators

Table 2 and Table 3 are clinical indicators of the voting algorithm in the chest X-ray images. A detailed evaluation of the majority algorithm was conducted from the clinical indicators. The voting algorithm correctly predicted chest radiograph images by 99.50%. The accuracy measure in this work denotes a total number of correctly identified images amongst all the images. The algorithm identified TB-infected images by 100% and healthy images 98.04%, this was dependent on the ability of the algorithm to distinguish features from the image. The principal component analysis with 98% variance ration helped the algorithm differentiate features by 94.83% which resulted in 100% ratio between correctly predicted images and all predicted images. The balance between healthy and TB-infected images was obtained by 95.15% from the mean accuracy and the sensitivity of the algorithm [17]. The accuracy of the algorithm prediction depends mainly on the quality of the input images [18]. Thus, the high accuracy of the algorithm is due to the linear nature of the data set. The linear kernel of the support vector machine produced a maximum margin between the two classes. On the other hand, the artificial neural network model was able to handle the data distribution by facilitating a lower learning rate. This implies that the optimal weights of the model were adjusted at a lower rate, and there were small changes in weights in response to errors. Another advantage is that the model had only one hidden layer; this reduced the overfitting by increasing the generalizability of the model on the test data. However, this was the major factor in the long prediction time of 92.98 seconds. The overall performance of the majority algorithm was confirmed by the high percentage (95.15%) of the f1 score.

In comparing of this work with previous work, S.I. Nafisah and G. Muhamad [19] developed an automatic TB detection from chest radiographs using deep learning models. Their best-performing convolutional neural network model (EfficientNetB3) achieved an accuracy of 99.10%. As such, the current work offers a slight improvement in accuracy. Although the works used different approaches, the slight improvement of the current work can be associated with the collaborative functionality of the algorithms as reflected in the F1 score. This collaborative functionality mainly depends on the number of models used in the voting process. This work used an odd number of models to avoid ties and overrule a vote of a model which may misclassify the input data.

Table 2. Clinical indicators of the voting algorithm on chest radiographs

Algorithm	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)
`SVM+DT+ANN`	99.50	100.00	94.83	100.00

Table 3. Continuation of clinical indicators of the voting algorithm on chest radiographs

Algorithm	PPV (%)	NPV (%)	F1 Score (%)	Time(s)
`SVM+DT+ANN`	100.00	98.04	95.15	92.98

4. Conclusions

Classification algorithms play an important role in the early screening of TB. This paper proposed a majority voting algorithm to classify chest radiograph images as healthy and infected with TB. The collaborative functionality of the models (support vector machine, decision tree, and artificial neural network) provides an accuracy of 99.50%. The voting algorithm offers an advantage to overrule missed predictions coming from one model. Another contribution to the accuracy of the algorithm was offered by the PCA, which enabled each individual model to detect sensitive information from the images while minimizing information loss. The use of the voting algorithm proposed in this work can be beneficial in medical facilities in providing second opinion to medical professionals by reducing the difficult task for clinical radiologists to classify abnormalities on chest radiograph images.

5. Acknowledgments

We acknowledge the Department of Science and Innovation (DSI) for funding this research.

6. References

[1] Rehman A, Khan A, Fatima G, Naz S. and Razzak I 2023 *Artif. Intell. Rev.* **2** 47
[2] Santos A D, Pereira B D B, Seixas J D, Mello F C Q and Kritski A L 2007 *Adv. Stat. Anal.* **4** 287
[3] Organization W H 2022 *WHO*
[4] Ndong Sima C A A, Smith D, Petersen D C, Schurz H, Uren C and Möller M 2023 *Immunogenetics* **75** 230
[5] Chakaya J, Petersen E, Nantanda R, Mungai B N, Migliori G B, Amanullah F, Lungu P, Ntoumi F, Kumarasamy N, Maeurer M and Zumla A 2022*Int. J. Infect. Dis.* **124** S29
[6] Mural R J, Adams M D, Miklos G L, Wides R and Romblad D L 2002 *sci.* **296** 1671
[7] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V and Vanderplas J 2011 *J. Mach. Learn. Res.* **12** 282
[8] Manaswi N K and Manaswi N K 2018 *Berkeley* **2** 43
[9] Van Der Walt S, Colbert S C and Varoquaux G 2011 *Comput. Sci. Eng.* **13** 30
[10] Ari N and Ustazhanov M 2014 *IEEE* **3** 6
[11] Quaranta L, Calefato F and Lanubile F 2021 *IEEE* **6** 554
[12] Kieu S T H, Bade A, Hijazi M H A and Kolivand H 2020 *J. Imaging* **6** 131
[13] Bhavsar H. and Panchal M H 2012. *Int. j. adv. res. comput.* **1** 189
[14] Liu W, Liu X and Chen X 2022 *SIAM J. Optim.* **32** 1957
[15] Rasamoelina A D, Adjailia F and Sinčák P *IEEE* **22** 286
[16] An L, Peng K, Yang X, Huang P, Luo Y, Feng P and Wei B 2022 *Sens.* **22** 821
[17] Nafisah S I and Muhammad G 2022 *Neural. Comput. Appl.* **21** 121
[18] Sanida T, Sideris A, Tsiktsiris D and Dasygenis M 2022 *Tech.* **10** 42
[19] Liang S, Ma J, Wang G, Shao J, Li J, Deng H, Wang C and Li W 2022 *Front. Med.* **9** 935080