IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Tuberculosis Images Classification Via Transfer Learning And K-Means

[1]Rahajaniaina Andriamasinoro, [2]Ratiarison Adolph Andriamanga
[1]Associate Professor, [2]Professor Emeritus
[1]Department of Mathematics Computer Science and applications,
[1]University of Toamasina, Toamasina, Madagascar

Abstract: Tuberculosis disease affects millions of people each year mainly in the developing country. Several researches were conducted to automatically diagnosis disease. Deep learning approach gave higher results in tuberculosis images classification. In the present work, we proposed a combination of transfer learning and unsupervised learning to clustering the Tuberculosis Chest Radiography imagines. Combination of MobileNetV3Small with lighter version of SSDLiteX auxiliary stage without top layers was used as feature extractor and unsupervised K-means clustering for classifying the features extracted. Two kinds of experiments were carried out and our model achieved an overall accuracy comparable to the other works. Additionally, our model could be run on the lightweight device as smartphone.

***Index Terms* -** Tuberculosis images classification, feature unsupervised learning, images clustering, Chest X-Ray.

I. INTRODUCTION

According to WHO, tuberculosis disease was the remaining the top killer among of people with HIV (WHO, 2024). Their report shown that in 2023, 10.8 million people worldwide fell ill with tuberculosis. The number of people who died reduce considerably compared to the last years. This situation was due to the evolution of research on the medical and others domains. Indeed, deep learning and transfer learning have been shown significative results in the last decade. Thus, in Xie et al. (2020) the authors used pre-trained models ResNet and EfficientNet for transfer learning to classify chest X-ray images of tuberculosis and they have an accuracy of 89.92% and 94.8% AUC. Others authors in Ayaz et al. (2021) performed a binary classification of CXR images into tuberculosis and normal using a Bayesian-based CNN (B-CNN) method. They tested their model with a combination of the Montgomery and Shenzhen datasets and they achieved an accuracy of 96.42%. As we know, these models based on the supervised learning. They used a pre-labelled dataset to perform the classification task. Or, in the real world, this situation is not always possible and medical dataset could be a big data. In fact, manual labelled is not possible. Furthermore, the use of the lightweight device is very popular in the developing country. Face to these situations, in the present work we suggested a lightweight model based on the combination of the transfer learning approach and unsupervised K-means clustering for classifying tuberculosis imagines. Comparative studies were conducted in order to demonstrate the performance of our model. This one is more less complex and could be run on lightweight device.

The rest of this paper is organized as follows: section II reviews related work of tuberculosis classification. Section III presents the material and methods. Results and discussion about our approach are discussed in section IV and section V conclude this paper.

II. RELATED WORK

This study focuses on the binary classification of pulmonary tuberculosis. Several kinds of works have already been carried out and continue to evolve.

In Nagireddy et al. (2024) the authors have been conducted a study to explore the effectiveness of SVM, Random Forest, Logistic Regression, Inception V3, and MobileNet in classifying chest X-ray images as normal or abnormal. They used OpenCV's histogram equalization and Contrast Limited Adaptive Histogram Equalization (CLAHE) image processing techniques to enhance image quality. Their experiment shows that the SVM provided the higher accuracy, precision and recall scores of 97% respectively than others models.

Osman et al. (2012) have been demonstrated a comparison of performance between clustering and thresholding techniques on images containing Tuberculosis bacilli from tissue using RGB images. During their study, they used K-means, moving K-means, fuzzy c-means as clustering algorithms and Otsu and iterative thresholding as adaptive thresholding ones. Each image was converted into C-Y color mode. The authors have been used the saturation component as input for each algorithm and their results show that K-means gave the best accuracy value of 99.49%.

The authors in Tawsifur et al. (2020) have been used the improved version of Canny edge detection algorithm on chest X-ray images. They have been combined the Shenzhen and Montgomery datasets together in order to have a total of 394 TB-infected lungs and 406 healthy lungs. They were discarded 12 normal images for balancing the dataset. This one was split into 90:10 ratios training/testing. After experiment, their system achieved an accuracy, sensitivity and specificity of 93.59%, 92.31% and 94.87% respectively.

In their work, Raharinirina et al. (2024) have been proposed a lightweight model based on the combination of MobileNetV3Small and an improved version of SSD Lite X (Hyeong-Ju Kang, 2023) to perform a binary classification of tuberculosis using chest X-ray images (CXR) from TB Chest Radiography Database. It contains 700 Tuberculosis images and 3500 normal images. Some pre-processing was necessary to ameliorate the image quality and to balance the dataset. Their model has been achieved the highest accuracy, recall, F1-Score and specificity of 100% respectively.

Michael et al. (2019) have been proposed VGG16 network to classify tuberculosis using 800 images obtained from a combination of Montgomery County dataset and the Shenzhen Hospital dataset. They have been used 75% of the dataset for training and the rest for testing. Their result shows an accuracy of 81.25%.

Hugo et al. (2020) proposed an hybrid technic to detect tuberculosis on lung radiographs. They have been used the SIFT and K-means algorithms. Their architecture was sub-divided into three stages: the first one was taking care of the pre-processing for re-scaling the images, the second one was used the SIFT algorithm for extracting the key points and performed the calculation of the keyCenters and the last one used these keyCenters for binary classification using K-means. They have tested the architecture with the Montgomery County X-ray set dataset with 300 lung radiographies. They have been applied k-Fold Cross-Validation with k=10 and achieving 90.3% accuracy.

Szilard et al. (2018) have been developed a system for detecting different lung problems from chest X-ray analysis. They used CLAHE technique for enhanced the image quality using an atlas-based lung segmentation algorithm based on model selection. The SIFT algorithm was used for extracting and selecting the best features in the lung region delineated. These best features have been used by SVM for classification into normal or tuberculosis. Their approach has been tested with two publicly available data collections: the Montgomery and the Shenzhen dataset and achieved an accuracy of 97.03%. Stefan Jaeger et. al. (2013) used a similar approach combining image processing technique and SVM classifies the computed feature vectors from chest X-ray images into normal or abnormal. The authors were tested their approach with two datasets: Montgomery country Dataset and The Shenzhen Dataset. For first dataset, they achieved an accuracy of 78.3% and 84% for the second one.

III. MATERIAL AND METHODS

Most of the previous work used supervised learning to classify tuberculosis imagines. These works gave better and efficient results with a labelled dataset. In contrast, labelled data is not always disponible. Additionally, it is possible that we will use an unlabelled big dataset. In this case, manual labelling data is not possible. To overcome this situation, in the present study, we proposed a combination of transfer learning and unsupervised learning to classify chest X-ray images in normal or tuberculosis. To achieve our goal, we used a similar approach as in (Hugo et al., 2020) and (Ryan et al., 2020) but instead of using SIFT algorithm or VGG16 network as feature extractor, we have been used a modified version of (Raharinirina et al.,2024), Fig.1 illustrate our approach.

3.1Dataset description

For this work, we used Tuberculosis (TB) Chest Radiography Database a publicly dataset downloaded from Kaggle web site. This dataset is a combination of three ones: NLM, Belarus and RSNA datasets. It is a pre-labelled dataset divided into normal and tuberculosis. It contains 700 TB images and 3500 normal images; the resolution of images is 512×512. To perform our study, we move all images from their original folder to a new one. This action provided us a dataset that contains all the images without labels.

3.2Data balancing and pre-processing

As mentioned in section 3.1, our dataset is imbalance and should be balanced to get higher performance of the model. The up-sampling technique were used to balance it. During balancing action, three transformations were applied: rotation with an angle of 30 degree, flip left right and flip top bottom. Then, some pre-processing must be applied to improve image quality. Thus, a contrast-limited adaptive histogram equalization (CLAHE) method combined with sharpening method were used for each image. Moreover, the weights of a neural network are trained using training images have a fixed size and have the same dimensions. The images in our dataset have 512×512 as resolution. To be compliance with the feature extractor, these ones must be resized into 224x224 pixels and normalized using MobilenetV3 built-in pre-processing function before passing through our feature extractor. Fig. 2 shows sample image before and after pre-processing from the dataset.

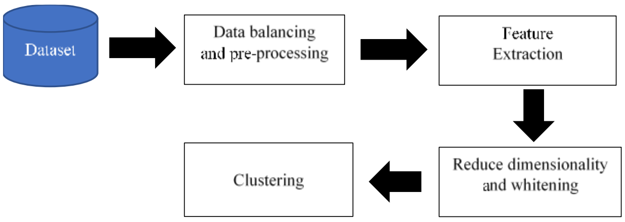

Figure 1: Our approach

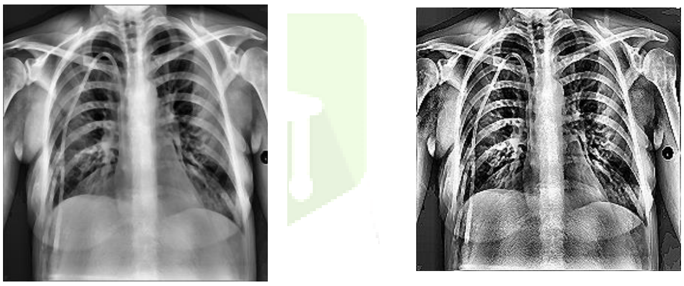

Figure 2: a) input chest X-ray image b) X-ray image after CLAHE
and sharpening

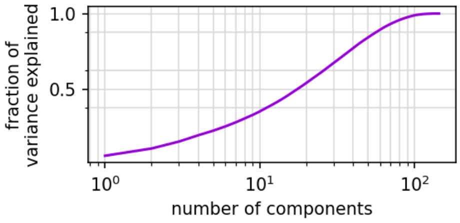

Figure 3: curve of the fraction of variance explained versus number of PCA components used for the features extracted

3.3 Proposed method

The aim of this study is to perform a lightweight model for classifying/clustering unlabelled chest x-ray images. After pre-processing, each image passed through the feature extractor. This one is (Raharinirina et al., 2024) without the top layer. In the other word, it is a combination of C4 and C5 layers of MobileNetV3Small (Andrew Howard, Mark Sandler, Grace Chu and al., 2019) and a lighter version of SSDLiteX auxiliary stage. We choose to use C4 and C5 layers because they gave the best features according to the state of-the-art. Then, principal component analysis (PCA) (Ian T. Jollife & Jorge Cadima, 2016) was applied with whitening to reduce the dimensionality of the data. Thus, the data are projected onto the n principal components that explain the highest fraction of variance. The choose of the value of n is a function of the variance-explained by each component (Ryan et al., 2020).

Using too many components is not interesting because it could introduce noise and have negative impacts on the quality of the classification. The use of whitening could minimize the correlation between the features and increase the performance of the classification. The Fig. 3 show the curve of the fraction of variance explained versus number of PCA components used for the features extracted. In our case, we use 34 components explains about 70% of the variance in the data. The rest capture only 30% of the data and it can discard for the classification task.

Next step is clustering, in which each image was assigned to a class label. For performing this step, K-means clustering (Hans-Hermann Bock, 2007) was applied with the k-means++ method. It is an unsupervised machine learning method that the principal goal is to group nearby points in feature space. The k-means++ method is one of different approaches to select the starting centroids. It helps K-means to perform good clustering performance and computational.

Additionally, K-means require an input value named k, which represents the clustering number. Several approach could use to obtain k value. In our case, its value was determined from characteristic of the original dataset: it contains two categories, normal and tuberculosis; this information gives us the number of the cluster i.e., the value of the k parameter.

IV. RESULTS AND DISCUSSION

The model was implemented on Intel(R) Core (TM) i7-1255U CPU, 10 cores, 12 threads @ 2.30GHz, and 24 Gb RAM using python, TensorFlow, NumPy, Matplotlib, Scikit-learn, Scikit-image, pickle, seaborn, pandas, OpenCV and Keras libraries.

Usual metrics for classification were used to evaluate the model. The accuracy, precision, recall and F1-score can be used to evaluate clustering algorithms while ground truth class labels are available. The accuracy gives the proportion of the correct prediction made by the model. Thus, this statistical measure computes the ratio for both true positives and true negatives to the true total number cases. Precision indicates the number of correct positive predictions from the total number of actual predictions classified by the model as positive. Recall in the other hand corresponds to the score of true positive predictions to the instances that actually belong to the positive class. The confusion matrix could be used also to evaluate the global performance of the model.

Table 1: comparative standard analysis results

Models	Accuracy	Precision	Recall	F1-score	Component number	Variance preserved	Support
VGG16	69.4%	81%	69%	66%	84	70%	7000
Our model	83.3%	83%	83%	83%	34	70%	7000

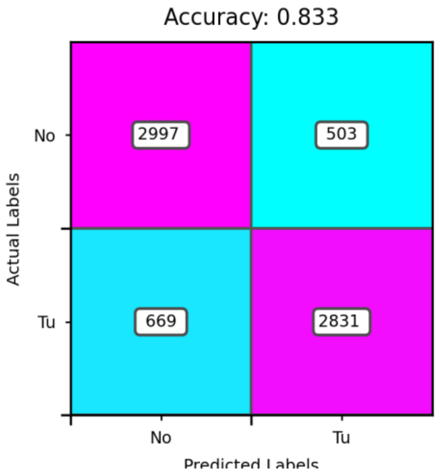

Figure 4: Confusion matrix of our model

Two experiments were carried out using VGG16 and our model: standard and predictive analysis.
In the standard analysis, after applied the reduction of dimension on the features, these ones were clustered using k-means clustering with k value equals 2, k-means++ initialization, and 500 different initialization steps. These actions were repeated 10 times with different initial random seeds in order to control the repeatability of the approach. For VGG16 architecture, we used the fc1 layer because it gave the best features according to the state of-the-art. The Table 1 shows the comparative result between VGG16 and our model. The Fig. 4 shows the confusion matrix, which it illustrates the global result of our model.

For the predictive experiment, we applied k-Fold Cross-Validation with k=10 on the dataset, i.e., 10-fold cross-validation to test the performance of our model and VGG16. Accuracy and F1-score were used as evaluation metrics for this experiment. The dataset was randomly divided into 10 equal groups. 9 groups have been used for the training set, were used to compute the PCA components and cluster centers using the standard analysis. The last one named validation set, was used to test the performance of the clusters on unsee data. The results of the analysis with cross-validation are shown in the Table 2.1 and Table 2.2.

As mentioned in section 3.3, our model is based on combination of MobileNetV3Small and a lighter version of SSDLiteX auxiliary stage, it is lightweight and less complex and could be run on a low-resource environment like a smartphone. It uses less computational resources to process images. To our knowledge, this is the first clustering model in the state of-the-art that could run on this kind of hardware.

Compared with others models, Table 1, Table 2.1 and Table 2.2 show that our model gave the best result than VGG16 in all evaluation metrics. in Melendez et al. (2015) the authors proposed an approach that combines Deep Learning with segmentation techniques to diagnose tuberculosis. They used a sample space of 800 radiographs and performed an average accuracy rate of 82.09 %. In the others words, they used small datasets compared to ours but they obtained results close to ours. This means that our model is acceptable and reliable to perform tuberculosis imagines classification.

Table 2.1: Predictive analysis results of VGG16

Train accuracy	Validation accuracy	F1-score
70 %	69 %	72%
70 %	67%	71%
70 %	70 %	74%
69 %	70 %	73%
69 %	70 %	73%
69 %	70 %	73%
70 %	70 %	73%
70 %	68%	72%
69 %	68%	71%
69 %	73%	75%

Average accuracy: 69.50% and Overall F1score: 72.64%

Table 2.2: Predictive analysis results of our model

Train accuracy	Validation accuracy	F1-score
83%	83%	83%
80%	78%	78%
83%	82%	82%
80%	80%	80%
82%	84%	84%
82%	83%	83%
83%	82%	82%
80%	80%	80%
80%	82%	82%
83%	82%	82%

Average accuracy: 81.83% and Overall F1score 81.87%

V. CONCLUSION AND FUTURE WORK

The present work was focused on tuberculosis imagines classification using the combination of transfer learning and unsupervised learning. We used PCA with whitening to reduce the dimensionality of the data and minimize the correlation between the features. This approach was tested with 7000 TB Chest Radiography images and was performed an overall accuracy comparable with the others works. According to its complexity, our model could be run on the lightweight device. In the future work, we will plane to test our approach with others datasets.

REFERENCES

[1] Andrew et al., 2019. Searching for MobileNetV3. arXiv:1905.02244v5 [cs.CV] 20 Nov 2019

[2] Ayaz et al., 2021. Ensemble learning based automatic detection of tuberculosis in chest X-ray images using hybrid feature descriptors. Phys Eng Sci Med 2021;44(1):183–94. https://doi.org/10.1007/s13246-020-00966-0

[3] Hans-Hermann Bock, 2007. \Clustering Methods: A History of k-Means Algorithms". In: Springer, Berlin, Heidelberg, 2007, 161-172. doi: 10.1007/978-3-540-73560-1_15.

[4] Hugo et al., 2020. Tuberculosis Detection Architecture with Image Processing using the SIFT and K-Means Algorithm, Computación y Sistemas, Vol. 24, No. 3, 2020, 989–997. https://doi.org/10.13053/CyS-24-3-3120

[5] Hyeong-Ju Kang, 2023. “SSDLiteX: Enhancing SSDLite for Small Object Detection”. Applied Science, 2023, 13, 12001. https://doi.org/10.3390/app132112001

[6] Ian T. Jollife and Jorge Cadima, 2016. Principal component analysis: A review and recent developments. Apr. 2016. doi: 10.1098/rsta.2015.0202.

[7] Melendez et al., 2015. On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis. IEEE Transactions on Medical Imaging, Vol. 35, No. 4, 1013–1024. https://doi.org/10.1109/TMI.2015.2505672.

[8] Michael et al., 2019. Pulmonary Tuberculosis Detection Using Deep Learning Convolutional Neural Networks, ICVIP 2019: 2019 the 3rd International Conference on Video and Image Processing, https://doi.org/10.1145/3376067.3376068

[9] Nagireddy et al., 2024. Enhancing tuberculosis detection through machine learning on chest x-ray scans. International Journal of Creative Research Thoughts (IJCRT), Volume 12, Issue 2 February 2024.

[10] Osman et al., 2012. Performance comparison of clustering and thresholding algorithms for tuberculosis bacilli segmentation. In: Proceedings of the CITS, 1–5 (2012)

[11] Raharinirina et al., 2024. Toward a lightweight model for pulmonary tuberculosis classification, International Journal of Advances in Engineering and Management (IJAEM) Volume 6, Issue 10 Oct. 2024, 308-32. https://doi.org/10.35629/5252-0610308321

[12] Ryan et al., 2020. Unsupervised machine learning via transfer learning and k-means clustering to classify materials image data. rXiv:2007.08361v1

[13] Stefan et. Al., 2013. Automatic Tuberculosis Screening Using Chest Radiographs. 2013 IEEE

[14] Szilard et al., 2018. Feature Selection for Automatic Tuberculosis Screening in Frontal Chest Radiographs. Med Syst (2018) 42: 146. https://doi.org/10.1007/s10916-018-0991-9

[15] Tawsifur et al., 2020. Reliable Tuberculosis Detection using Chest X-ray with Deep Learning, Segmentation and Visualization, IEEE Access, 2020, https://doi.org/10.1109/access.2020.3031384

[16] WHO 2024. The global Tuberculosis report 2024. https://www.kncvtbc.org//en/2024/10/

[17] Xie et al., 2020. “Computer aided system for the detection of multicategory pulmonary tuberculosis in radiographs”. J Health cEng. 2020; 2020:9205082. https://doi.org/10.1155/2020/9205082