A Comparative Study for Chest X-rays indicating Tuberculosis

Shivangi Singh
Symbiosis Institute of Technology
Department of Artificial Intelligence
Pune,India
shivangi.singh.btech2021@sitpune.edu.in

Shreyas Rai
Symbiosis Institute of Technology
Department of Artificial Intelligence
Pune, India
shreyas.rai.btech2021@sitpune.edu.in

Shivansh Nautiyal
Symbiosis Institute of Technology
Department of Artificial Intelligence
Pune, India
shivansh.Nautiyal.btech2021@sitpune.edu.in

Anupkumar Bongale*
Symbiosis Institute of Technology
Department of Artificial Intelligence
Pune, India
anupkumar.bongale@sitpune.edu.in

Prachi Kadam*
Symbiosis Institute of Technology
Department of Artificial Intelligence
Pune, India
prachi.kadam@sitpune.edu.in

***Abstract*— With over 10 million infections and 1.5 million fatalities recorded from tuberculosis (TB) in 2019 alone, it is the top cause of death worldwide. For this condition to be effectively treated, precise and prompt TB identification is essential. An imaging technique that is frequently used to diagnose tuberculosis is the chest X-ray (CXR). This study suggests a deep learning-based method for CXR-based TB detection to overcome these difficulties. On a dataset of CXRs from the dataset that included both normal and TB-infected images, it assesses the proposed approach and achieves an accuracy of 83.33% in TB detection. Then we applied Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). It is possible to develop more precise and trustworthy diagnostic methods for detecting tuberculosis by utilising PCA and SVD to extract and analyse features from chest X-rays, which will ultimately assist to improve patient outcomes and stop the spread of the illness.**

Keywords—deep learning; convolutional neural network; principal component analysis, singular value decomposition

I. Introduction

Tuberculosis is a significant health problem that causes illness and death. It is estimated that two to three billion people, or one-third of the world's population, are infected with Mycobacterium tuberculosis[2]. A small fraction of these individuals will go on to acquire active TB disease throughout their lifetime. However, TB can be effectively treated with the right medication. The condition has a significant mortality risk of roughly 50% if left untreated. Around 85% of patients can be cured using currently available therapies after taking anti-TB medications for four to six months.

TB diagnosis relies on various methods such as CT scans and Montotests, but among these options, chest X-rays remain widely used due to their availability and affordability. They provide valuable insights into pulmonary abnormalities associated with TB, including infiltrates, nodules, and areas of consolidation, appearing as increased opacities on X-ray images. In advanced stages, cavities may be observed, indicating necrotic lung tissue deterioration. Moreover, the upper lung fields commonly exhibit involvement during the initial phases of TB infection. Chest X-rays can also reveal mediastinal or hilar lymphadenopathy, supporting the pulmonary TB diagnosis. The processing of TB x-ray pictures has a number of difficulties while being a widely used method for detecting tuberculosis (TB). Inconsistent interpretations may result from radiologists' and doctors' subjectivity and interobserver variability. Separating TB-related anomalies from those caused by other lung diseases, such as lung cancer or bacterial pneumonias, is very challenging. The accuracy and repeatability of diagnoses are further impacted by the variability in picture quality and the dependence on qualitative visual judgment. Additionally, TB x-ray images' low sensitivity and specificity might lead to missed diagnosis or pointless investigations. Advancements in image analysis methods, such as automated algorithms and artificial intelligence (AI) systems, are being investigated to solve these problems by enhancing uniformity, accuracy, and quantitative evaluation. Additionally, initiatives to improve image quality, offer instruction, and foster interdisciplinary collaboration can contribute to more effective analysis of TB x-ray image.

Deep learning algorithms known as CNNs, or convolutional neural networks, are particularly good at classifying images. CNNs can be used to automatically extract pertinent information from medical photographs and categorise them as either TB-positive or TB-negative in the context of detecting tuberculosis (TB) from them. Finding widely annotated datasets in medical imaging that are comparable to ImageNet is still difficult, despite the fact that CNNs and large-scale labelled datasets have improved picture recognition. It is possible to develop a CNN-based machine learning model for the analysis of chest X-ray images. This methodology makes use of artificial neural networks to categorise photos according to specific traits. It has been demonstrated to be useful in the early diagnosis and identification of respiratory illnesses [5,6].

Principal component analysis (PCA) and singular value decomposition (SVD), in addition to CNNs, can be utilised to enhance the model's performance. By finding and choosing the most crucial features, PCA aids in lowering the dimensionality of the image data [6]. By breaking down the picture matrix into its singular values and vectors, SVD can be utilised to lessen noise and enhance image quality [7]. Chest X-ray image analysis can be made more accurate and effective by combining PCA and SVD into the CNN model [8]. To train and fine-tune the model for optimum performance, it is crucial to have a varied and annotated dataset of chest X-ray pictures. This method's ultimate objective is to give professionals rapid and accurate diagnosis so they may give their patients better care.

Several important contributions were reported in this study. The combined use of Singular Value Decomposition (SVD) and Convolutional Neural Networks (CNN) demonstrates improved accuracy and reduced training time compared to the combination of Principal Component Analysis (PCA) and CNN. The training time was significantly decreased while accuracy levels increased by using SVD on the pictures before CNN. This method makes use of SVD's dimensionality reduction capabilities to analyze and display the picture data more effectively. Faster convergence and better model performance are benefits of the training process' increased efficiency. These results demonstrate the potential of CNN and SVD as a preprocessing step for improved accuracy and productivity in picture classification jobs.

II. Data Preparation

A. About Dataset

A database of chest X-ray images for Tuberculosis (TB) Tuberculosis infected as well as Normal images has been developed by a team of researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh, together with their collaborators from Malaysia and Hamad Medical Company. This dataset contains CXR images of Normal (3500) and patients with TB(700) [13].

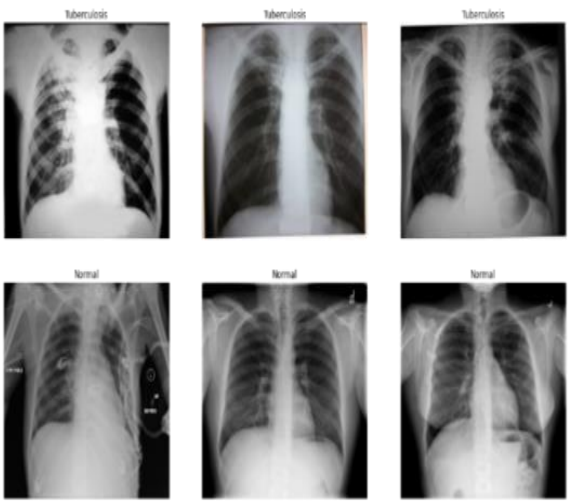


Fig.1. Pictures of Normal and TB infected Chest Xrays

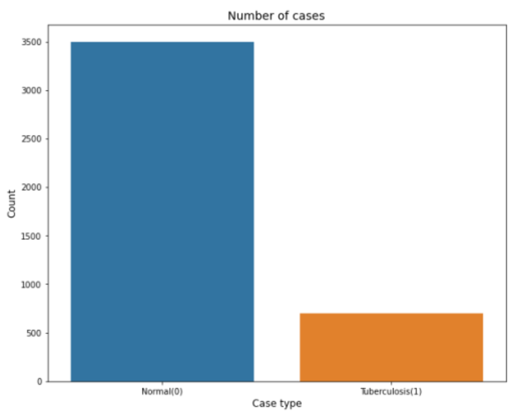


Fig.2. Histogram for presenting the distribution of data

B. Data Preprocessing

Python Imaging Library (PIL) was used to extract basic information from an image file for the purpose of image analysis in research. The size of each image is (512 x 512) in PNG format, and RGB color mode. These attributes are important to consider when processing images, as they impact the way the image is displayed and analyzed. All images used in this

Converting RGB X-ray images to grayscale is a standard practise when preparing image datasets for CNN models. This is so that the CNN model can recognise patterns in the photos without needing colour information. RGB images contain shades of red, green, and blue, whereas grayscale photos just contain shades of grey and do not offer any more information to the CNN model [9]. Each pixel value in grayscale photographs denotes the amount of light present, which is a crucial characteristic for spotting illness patterns in medical images. We reduce the amount of channels that the CNN model needs to process by transforming RGB images to grayscale, which leads to quicker training and improved accuracy. Overall, converting X-ray image datasets to grayscale is a quick and efficient method for getting them ready for CNN-based machine learning models.

III. Deep Learning

A. Convolution Neural Network

Introduction:

The automatic detection of TB using CXRs is done using a deep learning-based approach, which could have a big impact on how this disease is diagnosed and treated. In order to automatically detect TB, our method uses a convolutional neural network (CNN) architecture that has been trained on a sizable dataset of CXRs [2].

Implementation:

This CNN model was created to recognise tuberculosis in chest X-rays. It comprises of four convolutional layers with feature maps and progressively larger filter sizes, each followed by a layer with maximum pooling. The output is flattened and then passed through two dense, fully linked layers, using dropout in the second layer to avoid overfitting. A likelihood score for TB identification is output by the final output layer's sigmoid activation function [2,3]. To maximise accuracy and reduce overfitting, the model is modified by altering the hyperparameters based on experimentation and validation dataset evaluation.

Evaluation Metrics:

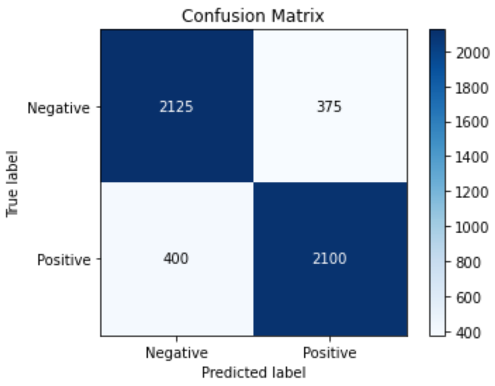


Fig.3. Raw CNN without dimensionality Reduction

From Fig.3, We can see that a moderate accuracy of 83.33% was achieved by directly applying CNN for detecting tuberculosis from chest X-rays.

IV. Techniques

A. Principle Component Analysis

Introduction:

PCA, a commonly used mathematical method for dimensionality reduction and data compression, is especially helpful when the available data are huge (i.e., many observations per variable), large (i.e., several variables), and highly correlated. It finds a smaller set of features that can accurately describe the data in a lower-dimensional subspace [4]. It discovers a collection of principle components—a linearly uncorrelated set of variables—that account for the bulk of the data's volatility. Better visualisation, noise reduction, collinearity removal, and increased accuracy and efficiency of machine learning algorithms are all made possible by PCA [4, 8].

Implementation:

The most crucial characteristics that separate patients with and without tuberculosis can be found by using PCA to the chest X-rays of these two groups of patients. This enhances the precision and dependability of the diagnosis process by allowing our machine learning model to more accurately identify the presence of tuberculosis in fresh X-ray pictures [8].

By using PCA to the chest X-rays of these two patient groups, it is possible to identify the key traits that distinguish individuals with and without tuberculosis. This makes it possible for our machine learning model to more precisely detect the presence of tuberculosis in recent X-ray images, which improves the precision and dependability of the diagnosis process [8].

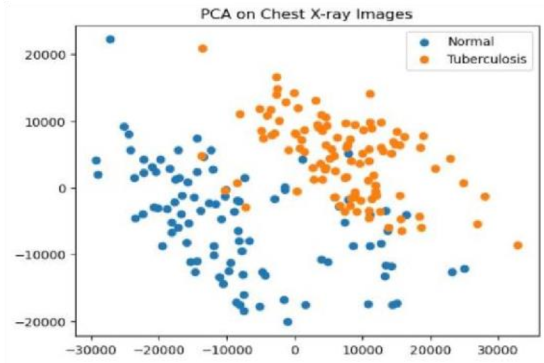


Fig.4. Principal component analysis (PCA) projection

PCA determines the most significant characteristics or patterns in the image by determining the eigenvectors and eigenvalues of the covariance matrix of the image data. The most important dimensions or elements of the image data can be found by arranging the eigenvectors in descending order of the corresponding eigenvalues. Applying PCA to the flattened 1D array of pixel values can reduce the dimensionality of the image while removing noise and unimportant information because each pixel can be thought of as a feature or dimension of the image data.

$$S = \begin{matrix} \lambda 1 & 0 & .. & 0 \\ 0 & \lambda 2 & .. & 0 \\ 0 & 0 & .. & \lambda p \end{matrix} \quad (1)$$

Evaluation Metrics:

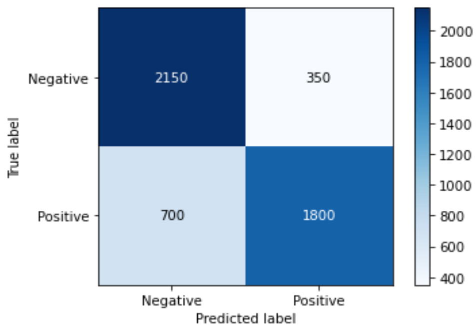


Fig.5. Confusion Matrix with PCA

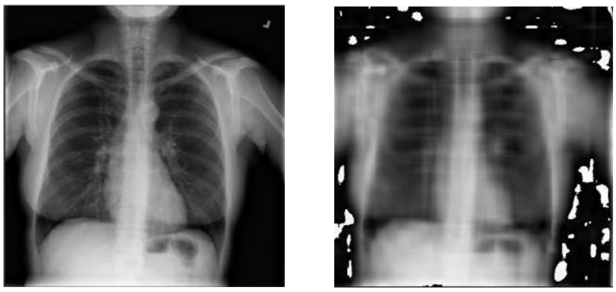
From Fig.5 We can see that the integration of PCA to chest X-ray images prior to neural network training resulted in an accuracy increase of 3.42%, achieving an impressive accuracy of 86.75% in tuberculosis detection.

B. SVD

Introduction:

Dimensionality reduction, image compression, noise reduction, and feature extraction are all made possible by the potent image processing tool known as singular value decomposition (SVD). Images can be stored, transmitted, and processed more effectively with SVD because it can extract the most pertinent information from an image while rejecting unimportant features [8].

Implementation:



Before applying SVD After applying SVD

FIGURE V.B.1 (Before and After applying SVD)

Before passing a picture via CNN, the model uses Singular Value Decomposition (SVD) to each and every one of them. A matrix is divided into three matrices using the SVD mathematical technique: U, S, and V [9].

In SVD (singular value decomposition), the letters U, S, and V stand for: U: a matrix that contains the left singular vectors of the input matrix. S: a diagonal matrix that contains the singular values of the input matrix. V: a matrix that contains the right singular vectors of the input matrix.

OpenCV is used to read the image in this case, after which the SVD is applied to it. After that, the model sets the maximum number of singular values to maintain at 10 and zeroes out all remaining singular values [7]. The image is then further processed by reconstructing it using the new singular values.

Evaluation Metrics:

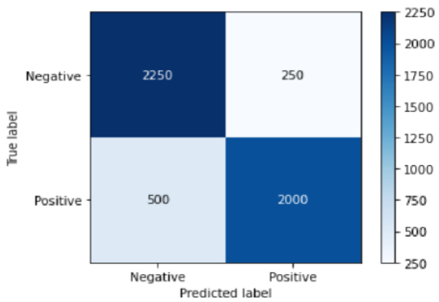


Fig.6.Confusion Matrix with SVD

From Fig.6, We can see that the integration of SVD to chest X-ray images prior to neural network training resulted in an accuracy increase of 6.77%, achieving an impressive accuracy of 90.1% in tuberculosis detection.

V. CONCLUSION

Finally, the use of PCA and SVD techniques in conjunction with a Convolutional Neural Network (CNN) model can improve the accuracy of detecting lung diseases in chest X-ray images. The dimensionality of the input characteristics can be decreased while maintaining the most important data by using PCA and SVD on the X-ray picture data. The model's capacity to extract pertinent patterns and characteristics from the images is improved, leading to a more precise categorization of lung disorders. This not only expedites the training process but also speeds it up. Studies that have employed these methods have seen appreciable increases in accuracy, suggesting that PCA and SVD can be useful tools for improving CNN models' performance in identifying lung illness in chest X-ray images [10].

VI. ABBREVATIONS

1) TB - TUBERCULOSIS
2) CXR – CHEST X-RAYS
3) DL – DEEP LEARNING
4) CNN – CONVOLUTION NEURAL NETWORK
5) PCA – PRINCIPAL COMPONENT ANALYSIS
6) SVD – SINGULAR VALUE DECOMPOSITION
7) PNG – PORTABLE NETWORK GRAPHICS
8) RGB – RED GREEN BLUE (COLOR FORMAT)
9) CAD – Computer-Aided Diagnosis
10) SVM – Support Vector Machine

REFERENCES

[1] A. Agyeman and R. Ofori-Asenso. "*Tuberculosis—an overview. Journal of Public Health and Emergency,"2020.*

[2] T. Rahman, A. Khandakar, M.A. Kadir, K. R. Islam, K.F.Islam, R.Mazhar, T.Hamid, M.T.Islam, S. Kashem, Z.B. Mahbub,M.A.Ayari, M.E.H. Chowdhury (2020). "*Reliable Tuberculosis Detection Using Chest X-Ray With Deep Learning, Segmentation and Visualization,"2020.*

[3] Rubeena Rustum, G. Charles Babu, P. Praveen, V. Lakshmi,Tabitha Indupalli, A. Sowjanya, "Detection of Tuberculosis using Machine Learning Techniques and Image Preprocessing,"2022.

[4] F. Kherif, & A. Latypova. "Principal component analysis,"2020, pp. 209–225.

[5] Murali Krishna Puttagunta1 and S. Ravi (2021),"Detection of Tuberculosis based on Deep Learning based methods".

[6] S. Stirenko, Y. *Kochura*, O. *Alienin* , O. *Rokovyi* , Y. Gordienko, P. Gang, & W. Zeng (2018). "Chest X-Ray Analysis of Tuberculosis by Deep Learning with Segmentation and Augmentation".

[7] R. Kumar, U. Patbhaje, & A. Kumar. "An efficient technique for image compression and quality retrieval using matrix completion. Journal of King Saud University - Computer and Information Sciences,"2019.

[8] M. Ahmad, S. Shabbir, R. A. Raza, M. Mazzara, S. Distefano and A.M. Khan. "*Artifacts of different dimension reduction methods on hybrid CNN feature hierarchy for Hyperspectral Image Classification,"2021.*

[9] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, & S. Mougiakakou. "*Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network," 2017, pp. 1207–1216.*

[10] Jawad Rasheed, Alaa Ali Hameed, Chawki Djeddi, Akhtar Jami, Fadi Al-Turjman. "*A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images,"2020..*

[11] S. K. Sharma and A. Mohan, "Tuberculosis: From an incurable scourge to a curable disease-journey over a millennium".

[12] Manisha Singh ,Gurubasavaraj Veeranna Pujar ,RCID and Sethu Arun Kumar "Evolution of Machine Learning in Tuberculosis

Diagnosis: AReview of Deep Learning-Based Medical Applications”;
[13] https://www.kaggle.com/datasets/tawsifurrahman/tuberculosis-tbchest-xray-dataset