Lung Problem Identification using Machine Learning targeting COVID-19

Muhammad Taha, Abdul Ahad, Syedah Nawal Munif, Muhammad Mustafa Usmani
NEDUET, taha4005680@cloud.neduet.edu.pk, hashmi4002872@cloud.neduet.edu.pk, ahad4000344@cloud.neduet.edu.pk, usmani4000445@cloud.neduet.edu.pk

Abstract - **The objective of this paper is to devise an efficient machine learning model to classify between COVID-19 and non-COVID-19 patients. We enlisted hundreds of features and extracted out the ones that had the most effect on the model performance. Our model uses a neural network that works on balanced as well as imbalanced datasets. We employ different techniques in order to successfully transform chest x-ray images into data suitable for processing. We also used the global features of the CT x-ray instead of focusing on some part of the image. The features that we use exist in both spatial and frequency domain which is one of the key factors on our model's success. This research and the following model can help hospitals triage resources on patients who need them most.**

Index Terms - COVID-19, coronavirus, medical image analysis, neural network, general image classification

INTRODUCTION

The most used method of testing for COVID is using a swab to extract a sample and then test that sample for the presence of virus. Another method is to take a blood test and look for the antibodies in the blood sample. Either of these methods take a couple of days or more depending on the priority. Chest X-ray (CXR) is a relatively cheap and accessible method for examining various lung problems. Hence why the CXR images are so readily available as completed datasets. According to a study, doctors can come to a conclusion by looking at a CXR for potential lung diseases with 95% accuracy [1]. That was our optimum result that we wanted to achieve with this project. We opted for a multi-class learner that could classify among a healthy chest, a COVID-19 infested chest and a pnuemonic chest. Our approach results in optimal accuracy and sensitive conclusions.

MATERIALS AND METHODS

Source Of Dataset

The dataset was taken from here: http://14.139.62.220/covid_19_models/

We took 1,950 samples of COVID-19, 416 of Pnuemonia and 192 CXRs of healthy lungs. The dataset is filtered on some of the following criterias: Must be adult, clarity of lung fluids, must be PA as it's usually preferred in terms of image quality and is more common so introducing unnecessary idiosyncrasies in a dataset may prove counterproductive in terms of extracting pattern from the data.

Dataset Transformation & Distribution

As the number of images was limited, we generated 25 different types of augmentations (Fig 1) through an open-source augmentation tool CLoDSA.

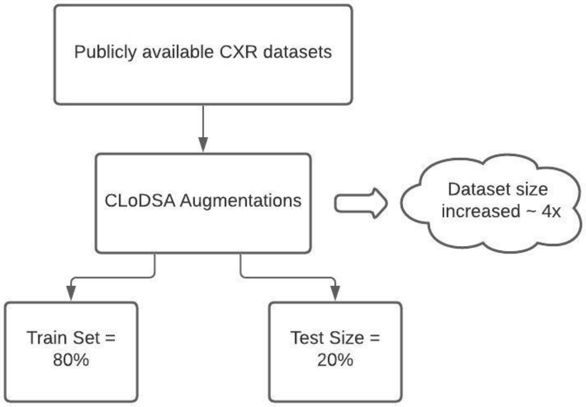


FIGURE 1: TRANSFORMATION AND DISTRIBUTION OF DATASET

For the training and testing of AI-based models, the original image dataset was divided into 75% training dataset (2,046 images) and 25% external validation dataset (512 images). We also split the dataset into 80% training and 20% internal validation dataset for hyperparameters optimization. In total, the dataset consisted of 2,558 images distributed unevenly in three classes.

Data Preprocessing Techniques

When looking at a CXR, it's a common practice to perform lesion segmentation so only areas that are relevant to the problem in the CXR is provided to the model. However, automated lesion segmentation is a complex machine learning project in and of itself so we used the global features of the entire CXR and presumed that our model would be able to extract the most distinguishing patterns. So, we only focused on normalizing the RGB pixel values to [0, 1] range and increasing the overall contrast of the CXR images.

Feature Extraction

One of the most important factors in a ML model's performance are the features used when training the model. Images in themselves are not native to data processing for machine learning. So, we need to find a way to transform these images into a format that is native. We perform statistical image analysis to characterize the location and the overall variability of the images. We used these fourteen histogram-based texture features and statistical identifiers[2]:

- Area
- Mean indicates general brightness of the image
- Standard Deviations signifies a measure of variability/contrast
- Skewness helps in observing image surfaces
- Kurtosis is inversely proportional to image noise
- Energy reveals how the gray levels are distributed
- Entropy helps with texture characterization
- Maximum
- Mean Absolute Deviation mainly used to signal noisy images
- Median
- Minimum
- Range
- Root Mean Square
- Uniformity

The identifiers were then used to extract features in 5 categories. Three from the spatial domain and two from the frequency domain. On a detailed study of CXRs, the most visually efficient way of telling the healthy and unhealthy lungs apart is looking into the texture of the important parts of the x-ray. So, the best way of making these images suitable for processing would be to perform texture analysis. Texture analysis is a very crucial task when performing image classification, segmentation and pattern recognition. It also directly affects the quality of features extracted and the model's performance.

We save the 14 identifiers defined above as they present a somewhat accurate description of texture in the spatial domain. However, it's sometimes better to evaluate images in the frequency domain as it may reveal patterns that normally are not observed just like sound is better analyzed in the frequency domain. Hence, we perform Fourier Transform on the image dataset.

Consider the FT of a two-dimensional sine wave on the left (Fig 2). The FT has the same dimensions in pixels as the original, and is entirely black except for a few bright pixels at the very centre. "If we zoom into the centre of the Fourier transform you can see there are exactly three pixels which are not black. One is the bright centre point, with coordinates (0,0), representing the contribution of the (0,0) wave to the image. The bright pixel on either side, with coordinates (1,0) and its reflection (-1,0), represents the contribution of the (1,0) wave (the sine wave in our original image). All the rest of the pixels in the Fourier transform are black, as the original image is exactly described using just the original (1,0) wave"[3].

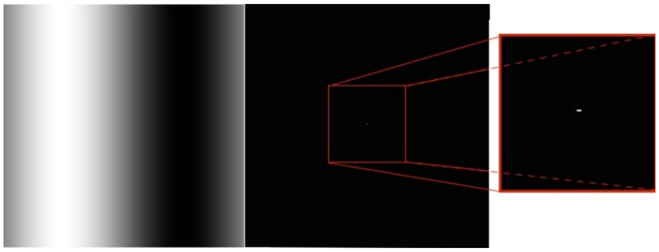


FIGURE 2: SIN(X) AND ITS FOURIER TRANSFORM

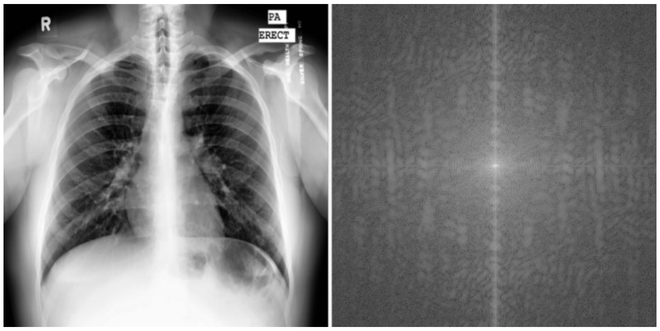


FIGURE 3: FOURIER TRANSFORM ON RIGHT

The image (Fig 3) on the right may look convoluted and complex but that's because digital photographs and pictures require many waves to represent them.

Rotating the image 45 degrees clockwise shows how crucial of a role the waves on these axis play in the final image (Fig 4). They also suggest the symmetry of an image.

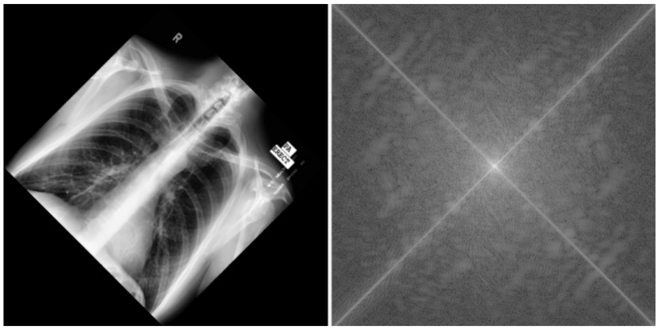


FIGURE 4: FIG 3 ROTATED 45 DEGREES

The third feature that we looked at was made by the way of applying Gray-level Co-occurence Matric (GLCM) to get the statistics on contrast and smoothness of the gray-level distribution of the image. We convoluted our dataset images with their respective GLCM kernel to get the output (Fig 5). GLCM helps to reflect the overall degree of correlation between pairs of pixels GLCM features belonging to the spatial domain and are based on the assumption that the texture information in an image is contained in the overall spatial relationship of the gray tones in the image.

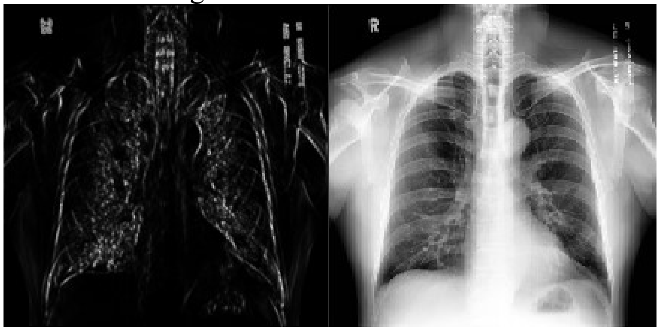


FIGURE 5: AN IMAGE WITH ITS GLCM ON THE LEFT

Much like the GLCM, GLDM is also a spatial domain feature and helps in calculating texture features like number non-uniformity, second moment and entropy. This matrix takes the form of a two-dimensional array Q, where Q(i,j) can be considered as frequency counts of grayness variation of a processed image. It has a similar meaning as histogram of an image[4].

The final method that we used to extract texture-based information was using wavelet transforms. Wavelet coefficients match different spatial frequencies in image under processing. Low and medium spatial frequencies usually match image content while high-frequency coefficients usually represent noise or texture areas. So, in wavelet domain you have an additional chance to distinguish image content and noise (Fig 6). It accurately analyzes the abrupt changes in the image that will localize means in time and frequency. Wavelets exist for finite duration and it has different size and shapes[5].

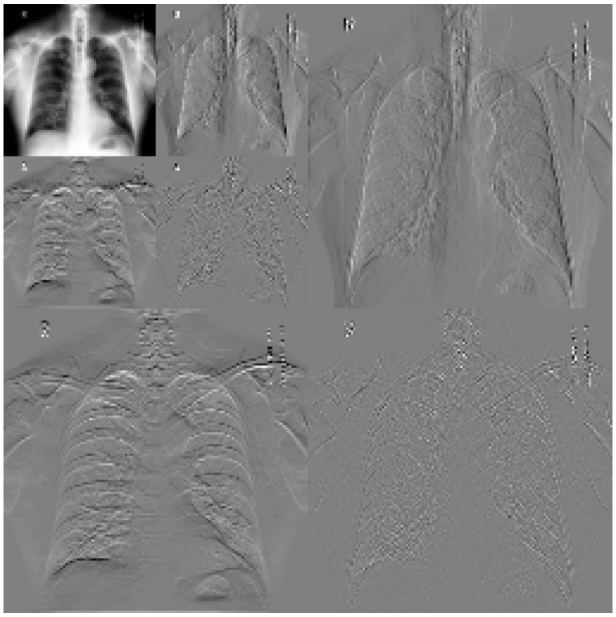


FIGURE 6: WAVELET TRANSFORMATION ANALYSIS

We made a feature pool of 252 columns and 2558 instances. This would result in a big model size. Hence, in an effort to reduce the number of features, we computed Pearson correlation coefficients (ρ) and plotted all the features with respect to ρ (Fig 7).

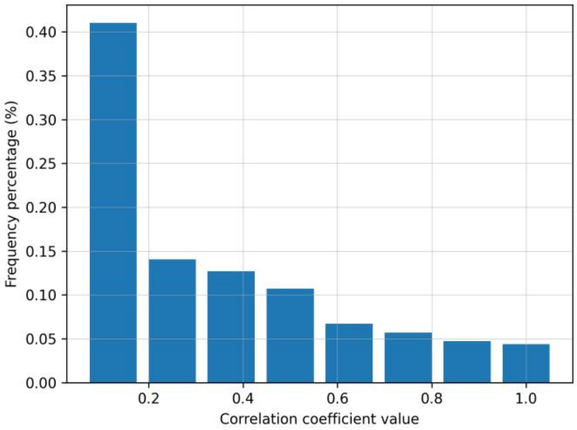


Figure 7: Pearson Correlation of features

Only around 25% of the features contributed to a $\rho > 0.4$ hence we used a dimensionality reduction technique through Principal Component Analysis which takes a linear n-dimension vector and transforms it into k-dimensions (usually $k < n$)

essentially reducing the number of features. We specified k = 64 as it is 25% of 252. We are now left with 64 features with decent correlation (Fig 8).

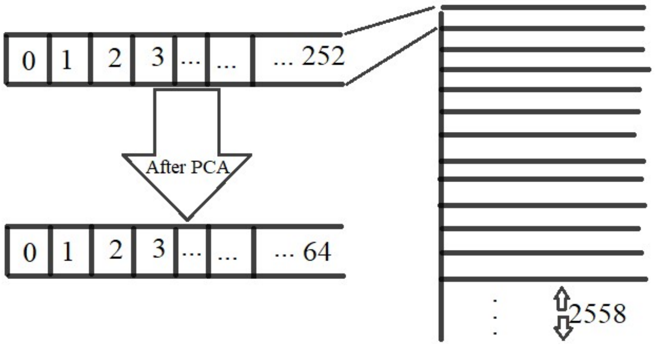


FIGURE 8: FEATURE POOL AFTER PCA

NEURAL NETWORK MODEL

We're using a multi-layer neural network for multi-label classification among the healthy, pneumonic and COVID-19 infested classes.

Multi-layer neural network (Fig 9) designed for the classification task including two hidden layers with 128 and 16 neurons respectively and a final classifier to classify the above mentioned cases.

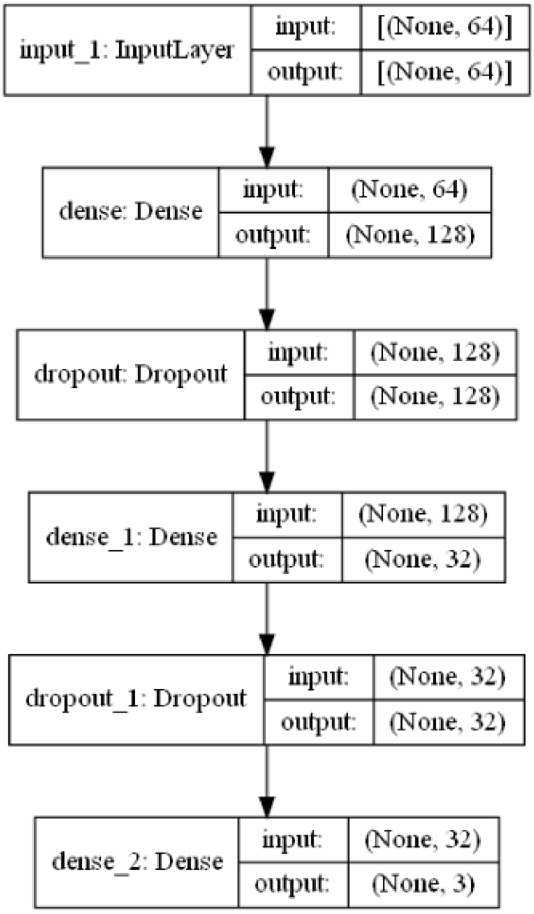


FIGURE 9: MODEL SUMMARY

The model performs considerably well with very few parameters i.e 12, 547. The model trains to about 26 epochs before stabilizing on its accuracy so the runtime is around 80-90 seconds on a modern machine for the entire model training.

RESULTS

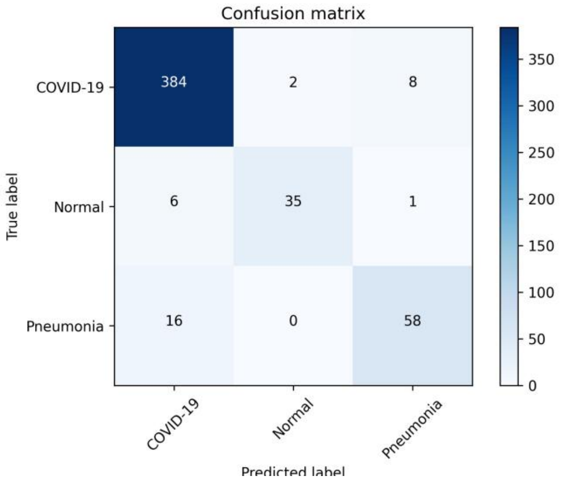
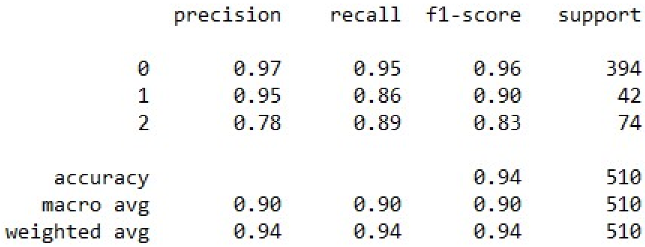


FIGURE 10: CONFUSION MATRIX

We immediately hit an accuracy of 82% in our first few epochs and stalled finally to ~94.3% and ~0.17 loss in both training and validation. Our precision comes out to be ~93.4% averaged across all three classes (Fig 10 & Fig 11).


```
              precision    recall  f1-score   support

           0       0.97      0.95      0.96       394
           1       0.95      0.86      0.90        42
           2       0.78      0.89      0.83        74

    accuracy                           0.94       510
   macro avg       0.90      0.90      0.90       510
weighted avg       0.94      0.94      0.94       510
```

FIGURE 11: CLASSIFICATION REPORT

We also plotted training loss to make sure our model was neither over-fitting nor under-fitting (Fig 12).

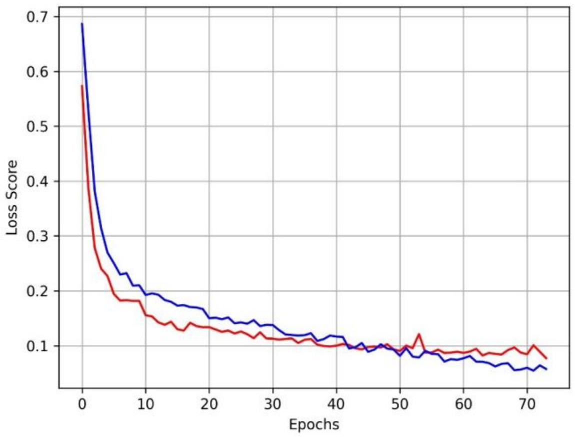


FIGURE 12: TRAINING LOSS

CONCLUSION

Our project serves as a proof of concept that by utilizing machine learning and AI we can come up with quick and acceptable solutions to a huge number of problems in many domains be it engineering, agriculture or as we showed it here: medical. We would like to assert that we believe with a few improvements our model can help with distribution of medical resources more effectively with faster and mass classifications.

REFERENCES

Following references were used to compile the code and this research paper.

[1] I Satia, S Bashagha, A Bibi, R Ahmed, S Mellor and F Zaman: "Assessing the accuracy and certainty in interpreting chest X-rays in the medical division" Aug 2013. https://dx.doi.org/10.7861%2Fclinmedicine.13-4-349

[2] Ravichandran, Dr. Purushothaman et al. "A study on Image Statistics and Image Features on Coding Performance of Medical Images." (2017). http://www.irdindia.in/journal_ijacect/pdf/vol5_iss1/1.pdf

[3] https://plus.maths.org/content/fourier-transforms-images

[4] NLGDM in MATLAB https://stackoverflow.com/questions/25019840/neighboring-gray-level-dependence-matrix-ngldm-in-matlab/25023396#25023396

[5] How will wavelet transforms be useful for image processing? https://www.researchgate.net/post/How_will_wavelet_transforms_be_useful_for_image_processing