Deep Learning for COVID-19 Diagnoses using Chest Imaging

William Phu
Computer Science and Molecular Biology
Massachusetts Institution of Technology
wphu@mit.edu

Lawrence Wong
Computer Science and Molecular Biology
Massachusetts Institution of Technology
lcwong@mit.edu

***Abstract*— Coronavirus disease 2019 (COVID-19) first emerged in the city of Wuhan in Hubei province and has become a global pandemic affecting millions around the globe. Due to the limited availability in traditional antibody testing, rapid and accurate diagnosis of respiratory diseases became more urgent in the midst of the pandemic. In this paper, we propose a machine learning tool based on chest radiographs (X-rays) that can aid radiologists or healthcare professionals in the diagnosis of COVID-19. X-rays is an imaging technique used to aid the diagnosis of many respiratory diseases, including tuberculosis (TB) and pneumonia. This image classification task is best accomplished by leveraging effective convolutional neural network (CNN) architectures. We aim to validate several methods, including logistic regression, K-nearest neighbors, and various CNN architectures, in the classification of posterior-anterior X-ray images of patients with different respiratory diseases. The experimental results show promising results in differentiating chest X-rays of COVID-19 from normal cases. Specifically, the ResNet architecture achieved a weighted accuracy of 99.0% (with a sensitivity of 97.0%, a specificity of 100.0%, and a precision of 100.0%) on the binary dataset. However, the performance dropped significantly as other respiratory images are mixed in to create the multiclass dataset. Saliency maps, filters, and activation visualization are used to interpret these techniques. While X-rays should not be used as first-line tests to diagnose COVID-19, we believe our findings can aid in decisions made by medical professionals.**

I. INTRODUCTION

In recent months, the COVID-19 outbreak became a pandemic that has swept the planet with over 3.5 million confirmed cases and over 250,000 deaths as of writing. COVID-19 is the disease caused by a novel virus known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). While most infected individuals have mild symptoms, serious complications can occur in a non-negligible subpopulation, including those who have preexisting conditions and those above 60 years of age. Complications like thrombotic and acute respiratory failure are common in critically ill patients with COVID-19 admitted to the intensive care units [15]. In addition, COVID-19 is considered to be a serious pandemic threat due to its ability to asymptotically spread during the incubation period. The basic reproductive ratio (R_0) of COVID-19 is reported to be between 1.4 to 6.49, with a mean of 3.28, exceeding that of severe acute respiratory syndrome (SARS) that plagued the world in 2002 [16]. R_0 measures the transmissibility of a virus since it represents the average number of new infections caused by an infectious individual. For transmissions to stop, the R_0 needs to be reduced to 1 through effective measures like immediate isolation of those who are infected. As a result, it is important to be able to quickly and accurately diagnose individuals to lower the transmission rate.

Current methods for laboratory testing fall into two categories: molecular tests that look for active infections and serology tests that look for previous infections. Molecular tests are based on reverse-transcription polymerase chain reaction (rRT-PCR) assays to detect active viral RNA in specimens taken from lower and upper respiratory tracks. While this form of testing can identify specific pathogens, the development is dependent upon understanding the genomic composition and expression patterns in the host during and after the infection [17]. Serology tests lack the ability to obtain immediate results since they are based on detecting antibodies. Antibodies specific to the virus are generally at measurable levels only after 7 days when symptoms become apparent [18]. Developing serological tests are also difficult due to the potential cross-reactivity of antibodies generated against other coronaviruses [18]. Moreover, both of these methods have dangerously high false-negative rates and long turnaround time, with some medical providers having to wait more than a week. Combined with logistical issues in the government and manufacturers, alternative solutions are desperately needed in the current situation.

Recent reports by radiologists suggest that chest imaging may help with diagnosing since infected indi-

viduals have X-ray patterns resembling those caused by pneumonia or acute respiratory distress syndrome [20]. COVID-19 patients commonly exhibit bilateral multi-focal consolidations that fill the pulmonary airspaces with fluid or other inflammation products in both lungs and small pleural effusions in the spaces around the lungs [20]. As such, recent publications have focused on leveraging deep learning models to identify patients who may be infected using chest X-rays and CT scans [1], [3], [4], [5], [6]. This project aims to build upon their models by comparing the performance of other non-neural network-based solutions to existing convolutional neural network architectures. In addition, we aim to expand the models to accurately classify other respiratory diseases using a multi-class dataset. We explored the VGGNet, ResNet, and DenseNet [21]. VGGNet extends upon AlexNet by stacking more layers and using smaller size filters to improving the performance of deep neural networks. ResNet is the first to use batch normalization and to popularize the idea of skipping connections to create residuals. Lastly, DenseNet expands upon the skipping connection idea of ResNet to help with the vanishing-gradient problem and strengthen feature propagation. We then use saliency maps, filters, and activation visualization to interpret our models. While models trained on the binary dataset showed promising results, the significant drop in performance when trained on the multi-class dataset needs to be further explored.

II. RESULTS & DISCUSSION

Based on receiver operating characteristic (ROC) curve, the best model for the binary dataset is ResNet with AUC-ROC of 1.0. The ROC curve is a probability curve that measures the performance at various classification thresholds. Fig.1. plots the true positive rate (also known as recall or sensitivity) against the false-positive rate. The area-under-the-curve (AUC) values are recorded in Table I for each classification model. Those values summarize the model's degree of separability between the positive and negative classes. An ideal model would have an AUC-ROC close 1.0 that predicts all disease examples as a positive class without making mistakes. However, there is a general trade-off between sensitivity and specificity as they are negatively correlated. Since the false-positive rate is calculated from specificity, true positive rate increases along with a false-positive rate in the ROC curve. AUC-ROC for models trained on multi-class dataset were not considered. This is because generalization to a multi-class dataset requires the one-vs-all method that uses a total of 20 curves to fully capture five labels.

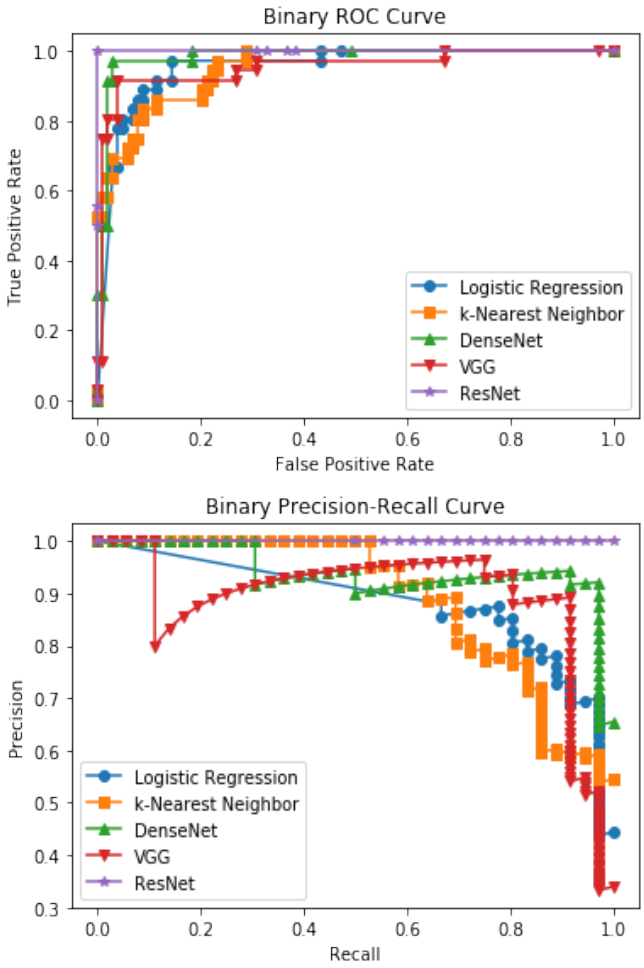

Fig. 1. The top shows the ROC curve while the bottom shows the precision-recall curve for the models trained on the binary dataset. Both graphs support the idea that ResNet has the best performance.

Based on precision-recall curves, the best model for the binary dataset is ResNet again with AUC-PR of 1.0. Precision-recall curves plot precision against the recall and is commonly used to measure prediction performance in imbalanced datasets. High AUC-PR indicates high precision and recall, meaning the model is accurately predicting all infected cases as positive. The model also has a high F1 score of 0.99. F1 score is the weighted harmonic mean of precision and recall and serves as another proxy to measure the model's accuracy. All results are summarized in Table I & II found on the last page.

A. Logistic Regression with L2 Regularization

The logistic regression model trained on the binary dataset achieved a mean 5-fold cross-validation score of

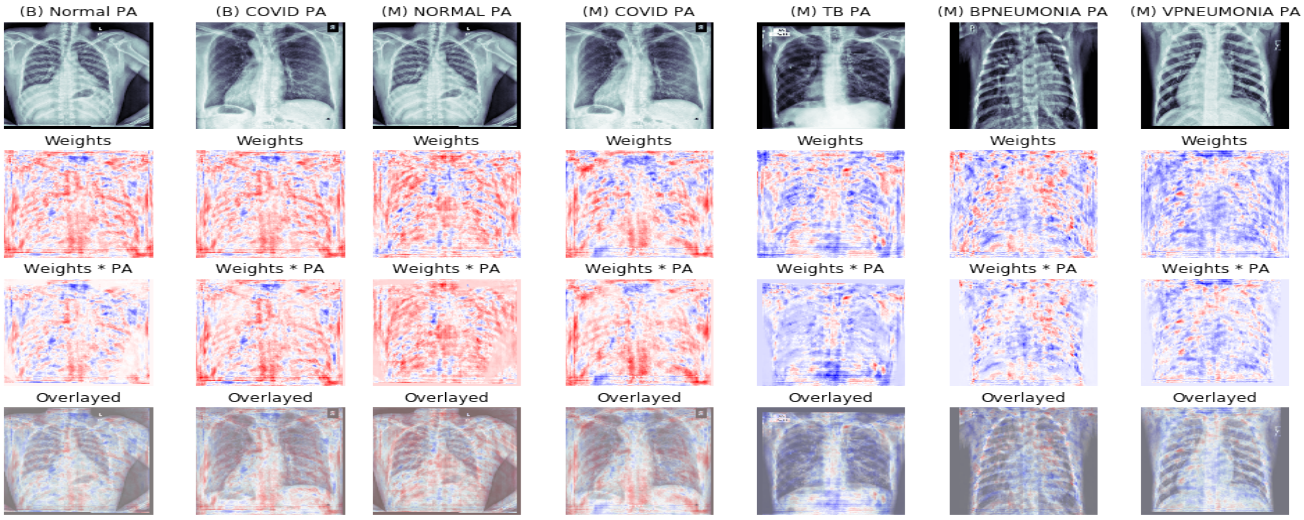

Fig. 2. Visualization of logistic regression weights for the binary dataset (left, B) and multi-class dataset (right, M). The figure consists of processed X-ray image (first row), reshaped weights from logistic regression (second row), values obtained from the dot product (third row), and overlayed images (fourth row). High activation regions are in red while low activation regions are in blue. Weights are prioritizing the thoracic regions for TB, bacterial pneumonia, and viral pneumonia. Strong activation on the border of the image brings and in the lumbar vertebrae region for the normal and COVID case brings concern about the validity.

0.81 and a weighted accuracy of 0.86 on the test split. This cross-validation score suggests that the model can generalize well to other independent datasets while maintaining high accuracy. There is a fine balance between precision (0.76) and recall (0.81) on the test split.

Logistic regression requires little computational resources, has no parameter to tune, and is highly interpretable. The model is trained on both binary and multi-class datasets. Fig. 2. visualizes the processed image and the coefficient for each pixel on the first and second row respectively. Coefficients with high values are indicated with red while low values are indicated with blue. The third-row shows the values after the dot product between the image and the weights. Finally, the fourth row overlays the third row with the first row to observe which area of the image is strongly activated. According to the figure, high-value weights are usually focused around the thoracic region as seen in the case of TB, bacterial pneumonia, and viral pneumonia. This is ideal since bacterial pneumonia typically exhibits focal lobar consolidation while viral exhibits more diffuse interstitial pattern in the thoracic region. However, it is concerning that high-value weights are also concentrated in the lower lumbar vertebrae and on the borders in the normal and COVID cases for both the binary and multi-class dataset. There have been no reports that respiratory diseases affect these areas.

While it does not require a linear relationship between the dependent and independent variables, this algorithm cannot handle non-linearities within the input dataset. It tends to fail tends and not perform well on datasets with large feature space. This is because features that are unrelated to the output and feature autocorrelations are introduced. This is important since pixel values tend to correlate with neighboring pixels. Flattening the input data also loses information about the relationship between neighboring pixels. The solver also had a very hard time trying to converge even at high iterations when training on high dimensional datasets.

The model's performance suffered significantly on the multi-class dataset. It uses a softmax function that biases towards a particular class when making predictions. The precision (.69) and recall (.66) for the COVID class is significantly higher than other classes despite having the lowest amount of testing examples. An improvement to this would be using the one-vs-rest method so that the predictions of each label does not correlate with one another. This idea is explored in the output layer of the neural networks. Nonetheless, logistic regression still serves as a good baseline model for comparison.

B. K Nearest Neighbors

The k-nearest neighbors model trained on the binary dataset achieved a mean 5-fold cross-validation score of 0.78 and a weighted accuracy of 0.87 on the test

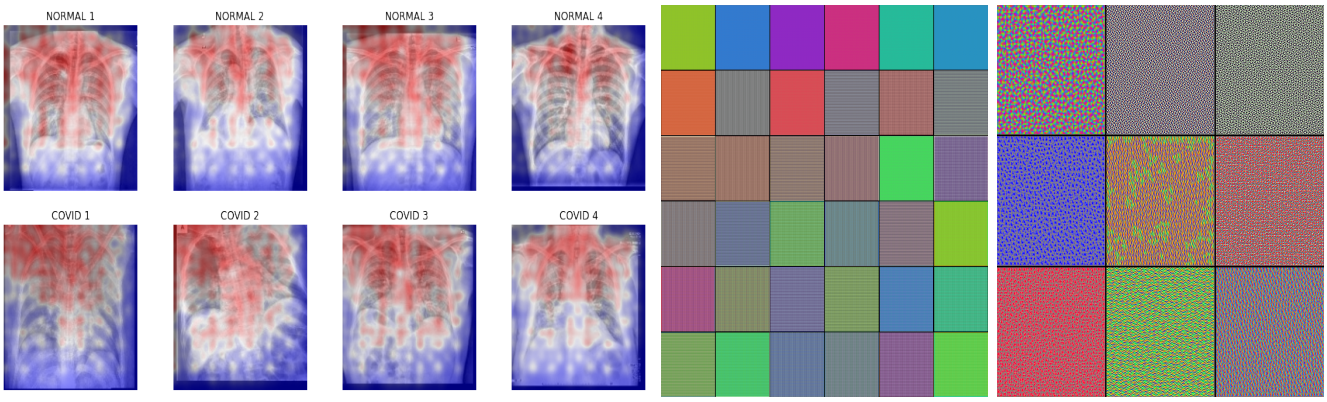

Fig. 3. ResNet saliency map visualizing input regions that cause the most change in the output (left). This image shows contributions from four normal X-ray images (top) and four COVID-19 X-ray images(bottom). Thoracic regions show the largest contributions (red) while abdominal regions show the smallest contributions (blue). ResNet convolutional layer activation visualization illustrates features that the model is detecting (right). On the second level of filters, some visible feature shapes begin to appear.

split. However, the precision (0.69) is much lower than the sensitivity (0.92). This suggests that the model is classifying many examples as positive but is not very accurate about it. This is likely due to the class imbalance issue. Since there are more training points for normal than COVID examples, it is more likely, by chance, that a new image will be closer to a normal data point than a COVID data point. Similar to the logistic regression model, cross-validation scores also show a decrease in performance in the case of the multi-class data. Higher weighted accuracy, precision, and recall might be due to a lucky split in data and is likely not significant. This result suggests that models based on global distances between examples are not so effective in multi-class classification of X-ray images.

C. Neural Network Models

Neural network models trained on the binary identification problem performed much better than the corresponding logistic regression and K-Nearest Neighbor models, with weighted accuracy scores above 0.9. Of the three models, the one that utilized ResNet for the basis of the network performed the best, with a weighted accuracy of 0.99. The specificity (1) and precision (1) indicates that the ResNet based model correctly identified all non-COVID samples, and the positive cases it identifies are in fact COIVD cases thus mitigating the potential risks of hospitalizing a healthy individual who might inadvertently become infected while under hospital care, as well as reducing fear and stress. The sensitivity (.97) is very high, although it indicates that the model is failing to classify a few COVID samples positively, which is worrying for its application as a diagnostic tool. The F1 score (.99) indicates that the model does well at balancing the precision as well as the sensitivity of the models. A perfect score of 1 would indicate that the model identifies all positive cases and only those cases. Overall, the other models followed similar trends, with specificity and precision being higher than sensitivity. This suggests that the models perform well at identifying healthy individuals as healthy individuals, but also identifying additional cases as positive and are getting it wrong.

In terms of raw performance for the binary case, the ResNet performs the best out of the three models, with the DenseNet ranking second, and the VGG a close third. This suggests that the problem can be improved by introducing more complicated or deeper models and that there still exists additional latent space that could be explored to further improve the results of a neural network approach. Additional steps such as finer grade hyper-parameter tuning may be able to improve the results of these models without needing to redesign the overall architecture.

The neural network models also outperformed the logistic regression and K-Nearest Neighbors models in multi-class classification, with all models scoring a weighted accuracy above 0.70. ResNet was the most successful of the three models, with a weighted accuracy of 0.77, and a weighted F1 score of 0.74. Again, like in the binary case, DenseNet performed second overall, and VGG came in third. The stark difference between the success of the binary case and the difficulties encountered with the multi-class case suggests that the models overall have a lot of work to do, inaccurately classifying a large fraction of the cases,

with some pathogenic cases labeled as normal, normal cases being incorrectly labeled as pathogenic, and one type of pathogenic cases labeled as another type.

While these models had lower accuracy compared to the binary case, this is to be expected due to the increased complexity of the multi-class identification problem. The lower success of the multi-class classifier suggests that a more complicated model may be necessary to be able to confidently predict from multiple classes of disease phenotypes. Additionally, it might be that the number of epochs necessary to fully train the model needs to be increased significantly, as there were several training instances where validation accuracy fell greatly, but then rose above the previous best nearly 10 or more epochs later.

One interesting note is that in both the binary and multi-class cases, ResNet consistently achieved the highest scores, with DenseNet second, and VGG third. As mentioned above, this affirms the suggestion that there is still information that has not been leveraged that deeper and more complicated architectures may be able to capture, since ResNet and DenseNet outperform VGG. However, it is not immediately clear why ResNet outperforms DenseNet in both scenarios. Perhaps the skip networks of residual networks are better suited for identifying very minute differences, and passing on only a select number of previous weights leads to some additional information that is useful when combined with the current layer, but having all of the information (through having all the layers connected) allows noise to interfere with the otherwise beneficial signal.

In order to understand what the neural network model was learning, we also visualized the predictions the model was using via saliency maps [22], which visualize which pixels contributed the most to prediction (Fig. 3). Saliency maps are a useful way to understand which features the model is emphasizing when it makes its predictions. This allows us to understand the decisions that the network is making. Saliency maps are visualized as heatmaps, with pixels with the highest activity getting the most intense values, and they are then overlaid on top of the image to identify features that the network is focusing on. The ResNet model shows high activation from areas of the image corresponding to the thoracic (upper) regions. Some of the lowest activations were found on the perimeter of the image, reaffirming that the model is learning from biologically relevant features rather than from features inherent to the processed image. We performed this analysis on both the binary and multi-class classification models, with similar saliency maps resulting. Saliency maps are especially useful in the multi-class case, as we can investigate whether or not the differences in activation between different lung disease phenotypes have a biological basis.

Additionally, the filters for the different layers of ResNet model were visualized for additional information. Visualizing the convolutional layers allows us to understand from another perspective what "features" the model is building, and understand how the model is interpreting the images. For example, the model might be developing features that detect solid masses in the lung tissue or maybe developing edge detectors (Fig. 3). When visualizing the convolutional layers, we saw that the initial convolutional layer is nothing more than rotations of linear lines, which suggests that there is some rotational information that the model is detecting, which is not necessarily relevant from a biological perspective. It may be possible to actually develop a convolutional filter that is rotationally invariant and thus reduce the number of filters used. Additional higher-level filters were investigated for both the binary and multi-class cases, but unfortunately due to computational constraints, they were infeasible to compute.

III. CONCLUSIONS

The successful results of the classification models, including Logistic Regression, K-Nearest Neighbors, as well as more sophisticated neural network approaches, indicate that machine learning-based models are well-suited as diagnostic tools for COVID-19. This task is best accomplished by neural network models, and even current architectures are up for the task of medical diagnosis without needing much work in additional design. Thus, it is more than possible to quickly and easily generate additional diagnostic tools that can be deployed. This also encourages the continued usage of chest x-ray scans as part of the diagnostic procedure when diagnosing patients with COVID-19, as well as the more general and wide-spread application of neural network models to clinical diagnosis that rely in any part on visual identification of features associated with a disease.

Some areas of improvement for evaluating these approaches would include the collection of additional images. Due to the novel and emerging nature of COVID-19, there are relatively fewer images publicly available. As time progresses, however, we expect additional datasets to be released, allowing the models to be trained on a larger, more quality set of images,

Fig. 4. Posterior-Anterior chest X-ray visualization of the normal case after each image processing step. CLAHE improves the contrast of the image while Kera's ImageDataGenerator rescales the image into appropriate inputs for training models.

thus hopefully improving the performance of these models. In addition, further hyper-parameter tuning could be performed. This step was limited due to time and computational constraints, but could easily improve the performance without needing more complicated architectures or additional data. Additionally, it is worth pursuing additional neural network models such as Inception, and NASNet to understand how these perform against the models tested in this paper. Given the promising results of the multi-class classification, it is also worth testing the model against an even larger class of respiratory disease to see whether these models have the capacity to generalize further.

IV. METHODS

A. Dataset Sources

Posterior-anterior chest X-rays of respiratory conditions in the study are collected from various sources. The tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA provides 80 normal and 58 tuberculosis X-ray images [11]. Shenzhen No.3 Hospital in Shenzhen, Guangdong providence, China also provides 326 normal and 336 tuberculosis X-ray images [11]. A COVID-19 image data collection project approved by the University of Montreal's Ethics Committee provides 80 normal and 142 COVID-19 X-ray images [13]. Kaggle's COVID-19 Patients Lungs X-ray images also provided 37 additional COVID-19 X-ray images [12]. Finally, a research paper about image-based deep learning provided over 400 bacterial pneumonia and 400 viral pneumonia X-ray from patients [10]. Bacterial pneumonia typically exhibits focal lobar consolidation while viral exhibits more diffuse interstitial patterns.

B. Image Processing

All images are first converted to JPEG format. Contrast Limited Adaptive Histogram Equalization

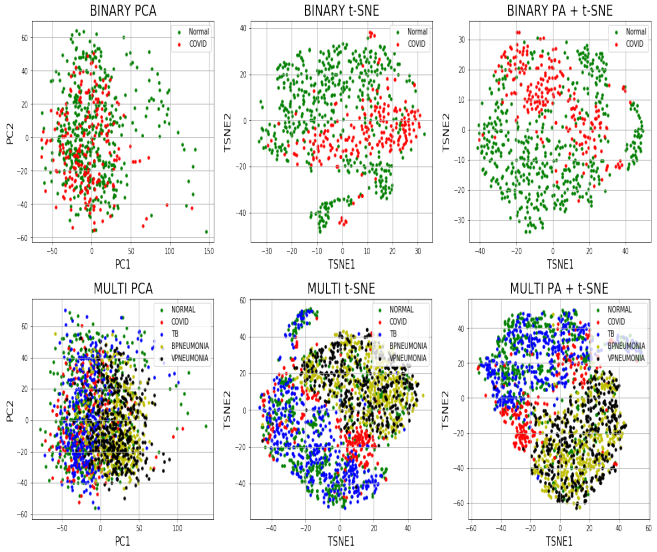

Fig. 5. PCA and t-SNE are used to visualize the flattened binary dataset with two labels and the flattened multi-class dataset with five labels. The third column shows the t-SNE of PCA reduced data to 50 dimensions. Clusters emerge in the t-SNE visualization.

(CLAHE) is then used to improve the contrast of the image by equalizing pixels within the image to cover a wide spectrum of values [7]. Adaptive histogram equalization additionally solves the issue of losing information due to over-brightness in global equalization by first sectioning the image into small tile blocks. CLAHE's output images are then fed into Kera's ImageDataGenerator [8]. Normalizing pixel to the range between [0, 1] restricts the weights range. Image ratio is locked into (224, 224), a common input size for CNN architectures. Pixels outside the image boundaries are filled with a constant black pixel. Images were allowed to be zoomed, sheared, and whitened based on parameters determined by training on a validation split. The outputs after each processing step are shown in Fig. 4.

Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (tSNE), are applied separately to transform the dataset into a 2D representation to visualize the structural patterns. PCA uses the correlation between dimensions to create new variables that retain the maximal amount of information about the original distribution. On the other hand, t-SNE is a probabilistic model that minimizes the Kullback–Leibler divergence between the pairwise similarity and lower-dimensional embedding distributions. A combination of both algorithms allows for faster t-SNE fitting time on a dataset that retains a high cumulative explained variation of the original dataset.

The lower-dimensional set is then colored with respect to each respiratory disease labels. The results are shown in Fig. 5, with binary dataset results on the top and multiclass dataset results at the bottom. No distinct clustering patterns were seen in the PCA. However, t-SNE projection method on the entire dataset and the PCA transformed dataset was able to find clustering patterns of the different respiratory conditions from X-ray images.

The image vectors are either flattened into a 1D array with shape (150528,) as inputs for baseline models or retained as a 3D array with shape (224, 224, 3) as inputs for the neural network. The dataset is divided based on sklearn's Train_Test_Split with a test size of 20% [9]. 20% of the training set is used as the validation set when fitting neural network models. All tested methods were five-fold cross-validated to measure how accurately the model will perform in practice on unknown datasets.

C. Logistic Regression with L2 Regularization

Logistic regression is a generalized linear model designed for binary classification but can be extended in the case of multi-class classifications. The algorithm finds the optimal weights θ, θ_0 that minimizes negative loss likelihood L_{nll} objective function.

$$J(\theta,\theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_{nll}(\sigma(\theta^T x^{(i)} + \theta_0), y^{(i)}) + \frac{\lambda}{2}||\theta||^2 \quad (1)$$

The feature x is a flattened image vector and the label y is an integer assigned to the corresponding respiratory disease case. Class weights are balanced inversely proportional to class frequencies in the input data. A ridge $L2$ regularization term is included to reduce model complexity and prevent overfitting. The solver algorithm used in the optimization is Limited-memory-Broyden–Fletcher–Goldfarb–Shanno (LBFGS). In the case of multi-class classification, a binary fit with corresponding weights is created for each label.

D. K Nearest Neighbors

Supervised K Nearest Neighbors classifies new data points by assigning it to the closest, distance-wise, cluster determined by a predefined number of training examples. The Euclidean distance function is used in the algorithm.

$$D(x_i, y_i) = (\sum_{i=1}^{k}(x_i - y_i)^2)^{1/2} \quad (2)$$

All points in each neighborhood are weighted by the inverse of their distance. This means that neighbors that are closer to the queried point have a greater influence than neighbors that are farther away. The Ball Tree algorithm is used as the solver to compute the nearest neighbors. K Nearest Neighbors can be easily generalized from binary to multi-class datasets.

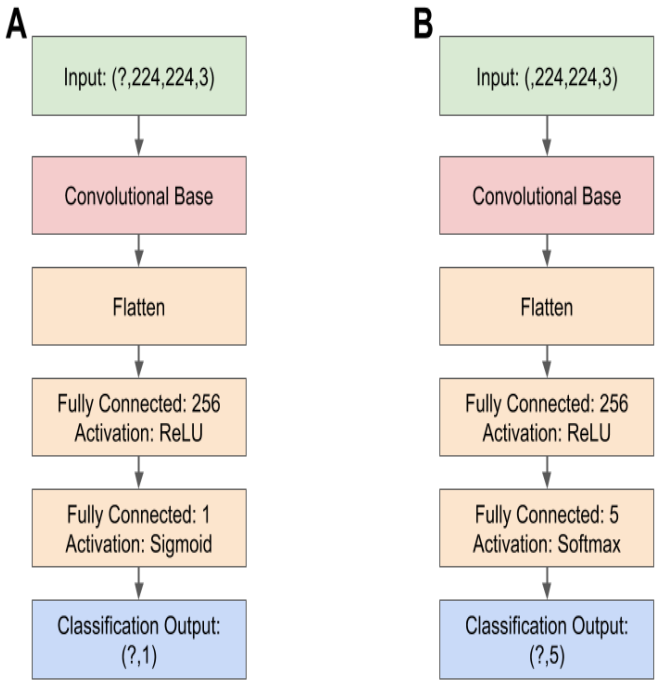

Fig. 6. Diagram of the neural network architecture. Three different convolutional bases were tested: ResNet, VGG, and DenseNet. A: Binary classification architecture, with sigmoid output. B: Multi-class classification architecture, with softmax output.

E. Convolutional Neural Networks

Three well known convolutional neural network architectures (DenseNet, ResNet and VGG) were integrated as bases (denoted as the convolutional base) into a complete convolutional neural network for binary (COVID-19 vs Normal) and multi-class (COVID-19, Viral pneumonia, Bacterial pneumonia, TB, and Normal) classification (Fig. 6). Image input shape is (224, 224, 3) for both architectures. The images are passed into the convolutional base, followed by flattening to 1-dimension, and then a fully connected layer with 256 neurons. In the binary classification case, the final hidden layer is a fully-connected layer with a single sigmoid output, denoting a value between 0 and 1 (Normal labeled as 0, COVID-19 labeled as 1). In the multi-class classification, the final hidden layer is a fully-connected layer with 5 neurons with a softmax activation, corresponding to the probability that the

image is of one of the five classes. The binary models were trained in batches of 20 for a total of 20 epochs while the multi-class models were trained in batches of 80 for a total of 30 epochs. Both models used a learning rate of 0.0005, keeping the epoch with the highest validation accuracy.

F. DenseNet Convolutional Neural Network

Dense neural networks achieve deeper network architectures by connecting each layer to one another in a feed-forward manner, such that all previous layers are used as inputs to the next layer. Dense neural networks attempt to mitigate the vanishing gradient problem and strengthen the propagation of weights. Three versions of DenseNet [23] with 121, 169, and 201 layers were tested.

G. ResNet Convolutional Neural Network

Residual neural networks aim to resolve the problem of neural-network degradation when stacking additional layers on top of deep networks by forcing layers to fit a residual mapping, which is related to the underlying mapping but is easier to optimize. Formally, if $H(x)$ is the underlying mapping, residual networks optimize the function $F(x)$, where

$$F(x) = H(x) - x$$

$$F(x) + x = H(x)$$

In feed-forward networks, this formula can be achieved through "skip networks", where layer l depends on the activations of $l - x$, where $x > 1$, thus skipping over the activations of the previous layer until it has matured. For this project, we built upon the ResNet_V2 model [14] with ImageNet weights as the basis for our residual network. We tested three versions of ResNet, with 50, 101, and 121 layers.

H. VGG Convolutional Neural Network

VGGNet [15] is a deep-convolutional neural network architectured by the Visual Geometry Group at the University of Oxford. VGGNet consists of multiple rounds of convolutional layers followed by max-pooling layers, and then a series of convolutional layers followed by a series of fully-connected layers which is then passed through a final softmax layer. For this project, we used both the 16 and 19 layer VGGNet architectures with ImageNet weights as the basis for separate neural networks.

ACKNOWLEDGMENT

The authors would like to thank Professors Gifford and Kellis as well as the teaching assistants, Sachit Saksena, Corban Swain, and Timothy Truong Jr for their guidance, support, and feedback. An immense amount of work and dedication went into making the class as wonderful as it was, and the authors learned new skills that will be invaluable in their future research. Their feedback was critical in narrowing the focus and scope of our project and helped highlight areas of confusion. Without the support of the staff, this class would not be half as interesting.

DIVISION OF LABOR

L.W and W.P collaborated equally on the collection of data. L.W processed and evaluated the quality of the image datasets, developed and tested the logistic regression and KNN models for binary and multi-class classification, and explored the activation of the logistic regression models and NN using saliency maps. W.P trained and tested the convolutional and recurrent neural networks for both binary classification and multi-class classification and explored filter activation. Both authors collaborated equally on writing the paper.

REFERENCES

[1] Abbas, Asmaa, et al. “Classification of COVID-19 in Chest X-Ray Images Using DeTraC Deep Convolutional Neural Network.” 2020, doi:10.1101/2020.03.30.20047456.

[2] Patil, Nandan L. “Current State and Predicting Future Scenario of Highly Infected Nations for COVID-19 Pandemic.” 2020, doi:10.1101/2020.03.28.20046235.

[3] Wang, Shuai, et al. “A Deep Learning Algorithm Using CT Images to Screen for Corona Virus Disease (COVID-19).” 2020, doi:10.1101/2020.02.14.20023028.

[4] Xiaowei Xu, et al. “Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia.” 2020, arXiv:2002.09334.

[5] Xu, Adrian Yijie. “Detecting COVID-19 Induced Pneumonia from Chest X-Rays with Transfer Learning.” Medium, Towards Data Science, 21 Mar. 2020.

[6] Zhou, Min, et al. “Improved Deep Learning Model for Differentiating Novel Coronavirus Pneumonia and Influenza Pneumonia.” 2020, doi:10.1101/2020.03.24.20043117.

[7] Laganière, R. (Robert), 1964-. OpenCV 2 Computer Vision Application Programming Cookbook : over 50 Recipes to Master This Library of Programming Functions for Real-Time Computer Vision. Birmingham, UK :Packt Pub., 2011.

[8] Chollet, Franọis, et al. "chollet2015keras," 2015. Available at: https://keras.io/

[9] Pedregosa, Fabian & Varoquaux, Gael & Gramfort, Alexandre & Michel, Vincent & Thirion, Bertrand & Grisel, Olivier & Blondel, Mathieu & Prettenhofer, Peter & Weiss, Ron & Dubourg, Vincent & Vanderplas, Jake & Passos, Alexandre & Cournapeau, David & Brucher, Matthieu & Perrot, Matthieu & Duchesnay, Edouard & Louppe, Gilles. (2012). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12.

[10] Kermany, Daniel Goldbaum, Michael Cai, Wenjia Valentim, Carolina Liang, Hui-Ying Baxter, Sally McKeown, Alex Yang, Ge Wu, Xiaokang Yan, Fangbing Dong, Justin Prasadha, Made Pei, Jacqueline Ting, Magdalena Zhu, Jie Li, Christina Hewett, Sierra Dong, Jason Ziyar, Ian Zhang, Kang. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 172. 1122-1131.e9. 10.1016/j.cell.2018.02.010.

[11] Jaeger, Stefan et al. "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases." Quantitative imaging in medicine and surgery vol. 4,6 (2014): 475-7. doi:10.3978/j.issn.2223-4292.2014.11.20

[12] Sajid, Nabeel. "COVID-19 Patients Lungs X Ray Images 10000." Kaggle, 23 Mar. 2020, www.kaggle.com/nabeelsajid917/covid-19-x-ray-10000-images.

[13] Joseph Paul Cohen and Paul Morrison and Lan Dao COVID-19 image data collection, arXiv:2003.11597, 2020 https://github.com/ieee8023/covid-chestxray-dataset

[14] Zhang, et al. "Identity Mappings in Deep Residual Networks." ArXiv.org, 25 July 2016, arxiv.org/abs/1603.05027.

[15] Klok, F.a., et al. "Confirmation of the High Cumulative Incidence of Thrombotic Complications in Critically Ill ICU Patients with COVID-19: An Updated Analysis." Thrombosis Research, 2020, doi:10.1016/j.thromres.2020.04.041.

[16] Liu, Ying et al. "The reproductive number of COVID-19 is higher compared to SARS coronavirus." Journal of travel medicine vol. 27,2 (2020): taaa021. doi:10.1093/jtm/taaa021

[17] Udugama B, Kadhiresan P, Kozlowski HN, et al. Diagnosing COVID-19: The Disease and Tools for Detection. ACS Nano. 2020;14(4):3822-3835. doi:10.1021/acsnano.0c02624

[18] Lv H.; Wu N. C.; et al.Cross-Reactive Antibody Response between SARS-CoV-2 and SARS-CoV Infections. bioRxiv, March 17, 2020. 10.1101/2020.03.15.993097 (accessed on March 20, 2020).

[19] Klimpel GR. Immune Defenses. In: Baron S, editor. Medical Microbiology. 4th edition. Galveston (TX): University of Texas Medical Branch at Galveston; 1996. Chapter 50. Available from: https://www.ncbi.nlm.nih.gov/books/NBK8423/

[20] Zu ZY Jiang MD, Xu PP, et al. Coronavirus Disease 2019 (COVID-19): A Perspective from China. Radiology. 2020;200490. doi:10.1148/radiol.2020200490

[21] Huang, Gao, et al. "Densely Connected Convolutional Networks." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, doi:10.1109/cvpr.2017.243.

[22] Simonyan, et al. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." ArXiv.org, 19 Apr. 2014, arxiv.org/abs/1312.6034.

[23] Huang, et al. "Densely Connected Convolutional Networks." ArXiv.org, 28 Jan. 2018, arxiv.org/abs/1608.06993.

TABLE I

BINARY MODEL SUMMARY STATISTICS

Model	CV	Weighted Accuracy	Precision	Sensitivity	Specificity	AUC-ROC	AUC-PR	F1 Score
Logistic Regression	0.81	0.86	0.76	0.81	0.91	0.95	0.88	0.89
K-Nearest Neighbor	0.78	0.87	0.69	0.92	0.86	0.95	0.89	0.88
DenseNet	0.94	0.97	0.90	0.97	0.96	0.98	0.94	0.96
VGG	0.94	0.93	0.89	0.89	0.96	0.95	0.89	0.94
ResNet	0.97	0.99	1.00	0.97	1.00	1.00	1.00	0.99

TABLE II

MULTI-CLASS MODEL SUMMARY STATISTICS

Model	CV	Weighted Accuracy	Weighted Precision	Weighted Recall	Weighted F1 Score
Logistic Regression	0.62	0.58	0.58	0.58	0.58
K-Nearest Neighbor	0.58	0.62	0.63	0.62	0.62
DenseNet	0.70	0.71	0.69	0.69	0.69
VGG	0.70	0.70	0.70	0.68	0.71
ResNet	0.73	0.77	0.76	0.73	0.74