Gina McFarland
COMP 4449 - Midterm
Winter 2022

Executive Summary

This project investigates how machine learning models perform on diagnostic x-ray imaging classification. 602 lung x-ray images were labeled normal, COVID-19, or pneumonia. After imaging preprocessing steps, including resizing, Gaussian blurring, and Histogram of Oriented Gradients (HOG), four different machine learning models trained on the data to create classification predictions. Support Vector Classifier (SVC) performed the best on the data with an accuracy of 93%. Random Forest and Multilayer Perceptron Neural Networks (MLP) achieved accuracy scores of 90% and 89% respectively. K Nearest Neighbor (KNN) performed the least well at 82%. These results were compared to a study in which radiologists diagnosed lung x-ray images of COVID-19 and pneumonia patients. The radiologists achieved an average accuracy of 74%. These results show that machine learning models have the potential to significantly enhance the field of radiology by aiding with diagnostic imaging.

Research Question

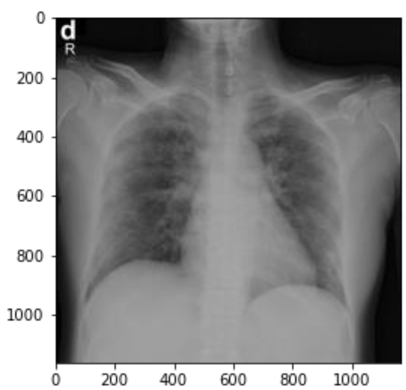
This project analyzes three sets of lung x-ray images and classifies whether the patient is healthy, has pneumonia, or has COVID-19. Specifically, the research question was to determine which of the machine learning algorithms tested, K Nearest Neighbors (KNN), Support Vector Machine Classifier (SVC), Random Forest, or Multilayer Perceptron Neural Network (MLP), can most accurately classify the patients' diagnosis of healthy, pneumonia, or COVID-19 based on their lung x-rays.

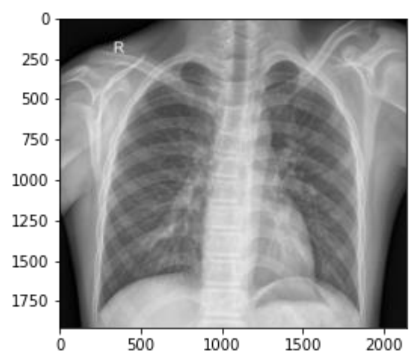
Using machine learning for diagnostic imaging has the potential to be disruptive in the field of radiology. While machine learning is unlikely to replace radiologists in the next 20 years due to technological, regulatory, and legal factors, it can enhance diagnosis accuracy for radiologists with an understanding of data science. (Chan, 2018) There are a number of challenges for accurate radiological diagnosis. Machine learning processing has the promise to alleviate some of these issues including excessive workload, visual and mental fatigue, staff inexperience, inadequate equipment, and poor lighting. (Brady, 2017) However, there are other technical problems that will make accurate diagnoses challenging even with data science techniques. These continued issues include the imaging protocol used, use of appropriate contrast in patients, and differences in patients' bodies. (Brady, 2017)

While the accuracy of radiologists varies significantly, in one study, radiologists had an average 74% accuracy rate distinguishing between COVID-19 compared to non-COVID-19 viral pneumonia. (Bai, 2020) The results of this analysis will be compared to this baseline.

Data

The data consisted of 602 lung x-rays, which were in folders labeled normal, pneumonia, and COVID-19, downloaded from Mendeley (https://data.mendeley.com/datasets/fvk7h5dg2p/1). (Shams, 2020) The normal and pneumonia sets had approximately the same number of images, with 234 normal lungs and 220 COVID-19 lungs. However, there were only 148 pneumonia lung images. The approximately 40% difference in the number of pneumonia lungs may cause an issue with the classification accuracy.





COVID-19 Normal Pneumonia

The images were in jpg, jpeg, and png file formats. The images above are shown prior to any processing, in their original file format. The images were read in using the skimage package which converted each image to a matrix with each pixel denoted as a number. The number indicates the image intensity.

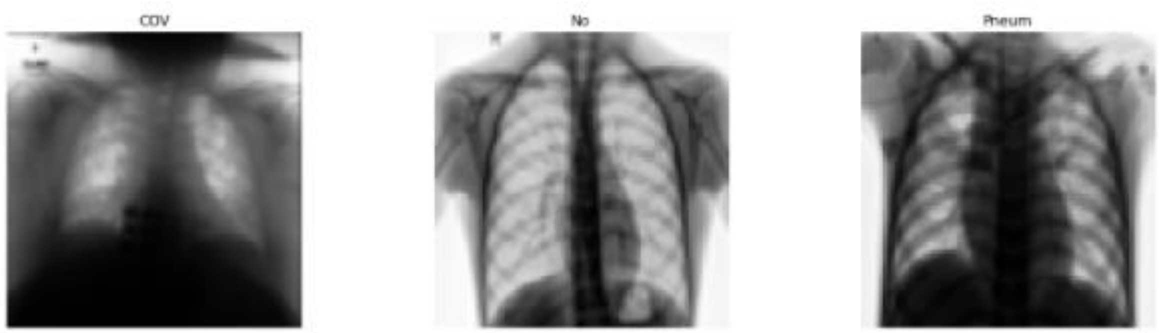
```
In [14]:  data['data'][0]

Out[14]: array([[0.33357475, 0.32585059, 0.35658545, ..., 0.48458416, 0.49176623,
                 0.6440686 ],
                [0.06627061, 0.07867302, 0.10656704, ..., 0.41650295, 0.42197002,
                 0.59504799],
                [0.01515436, 0.0243757 , 0.02660981, ..., 0.37107427, 0.37280089,
                 0.5591654 ],
                ...,
                [0.95162274, 0.95069136, 0.9570466 , ..., 0.8690419 , 0.83609848,
                 0.85974611],
                [0.95415969, 0.9528574 , 0.9596131 , ..., 0.88547885, 0.85449823,
                 0.87661561],
                [0.95606388, 0.95539292, 0.96233607, ..., 0.9012843 , 0.87240793,
                 0.89215804]])
```

The first image as stored in the dataset

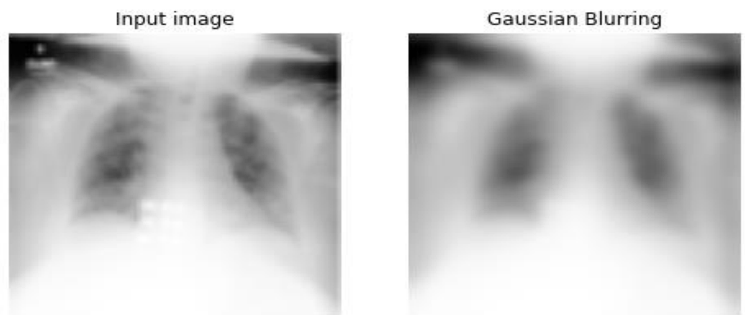
An analysis of the pixel intensity showed that the minimum intensity was 0.0, the maximum intensity was 1.0 and the average intensity was 0.5.

After reading in the images, the images were processed in the standard manner for image analysis preparation. This includes resizing the images to ensure standardization and removing image noise. (Canuma, 2020) To ensure data comparability, each image was resized to 80 x 80 pixels and converted to grayscale so that the images were stored in a two-dimensional array. Two of the datasets were already in grayscale and one was stored with a multidimensional channel containing color information. (Image analysis in Python)

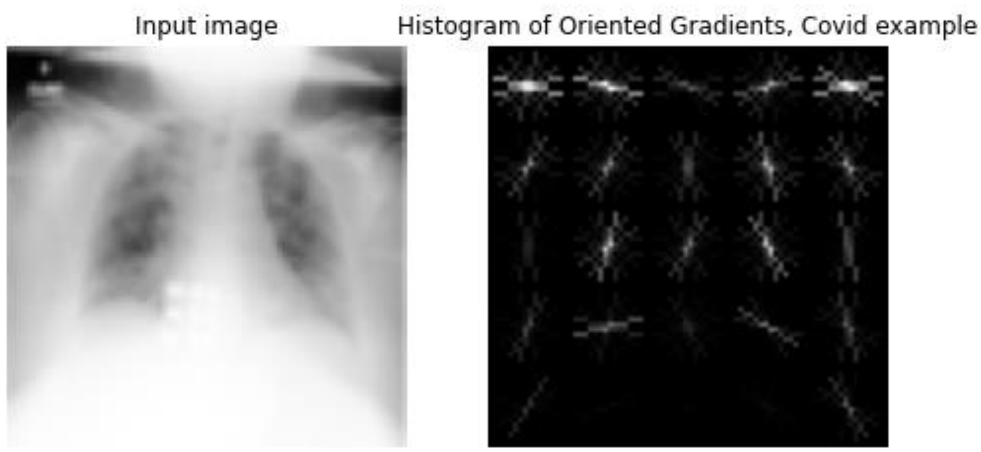


Resized, gray-scaled images

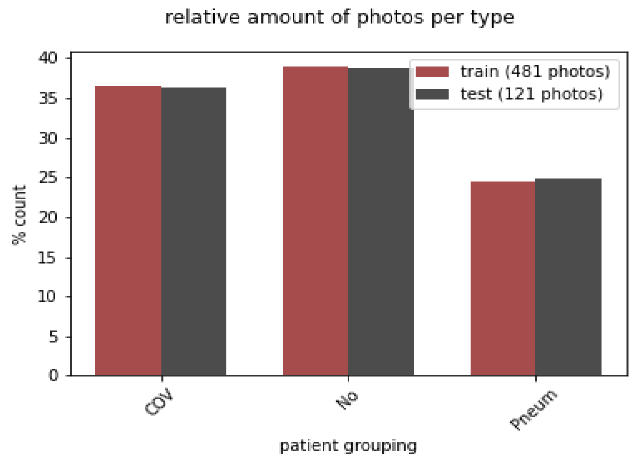
The images above are the stored images. Each picture is an equivalent size, and while the images are degraded from the originals, they still bear a strong resemblance. Next, image noise was removed with a process called Gaussian blurring. Gaussian blurring averages the pixel location but adds a higher weight to the center pixels. The concept is that the average variance of noise is smaller when considering the variance of a single pixel. There must be a fine tuning of the amount of blurring in order to remove noise, but not to remove image gradients. (Collins)



As a final data preparation step, a process called Histogram of Oriented Gradients (HOG) was used to feature engineer the data and simplify the image to its most important aspects. The method divides the image into sections, identifies edge pixels and extracts their gradient and orientation. (2020)



Post-processing, the data was split into 80% training and 20% test with the sklearn train_test_split method. Since the data was read in by folder, the images were sequentially aligned with COVID-19 patients first, then healthy patients, then patients with pneumonia. The flag shuffle was set to True to distribute the data among the test and training sets. Additionally, the flag stratify was set to ‘y’ in order to create even patient groupings in the training and test datasets. This helped to mitigate the imbalance in the pneumonia data set by including the same proportion of examples to the model in training and test.



Research

After the images were processed for analysis, the data was prepared to answer the research question: Which machine learning algorithm: K Nearest Neighbors (KNN), Support Vector Classifier (SVC), Random Forest, or Multilayer Perceptron Neural Network (MLP), can most accurately classify the patients' diagnosis of healthy, pneumonia, or COVID-19 based on their lung x-rays.

The first machine learning model used was K-Nearest Neighbor (KNN) Classifier. The KNN algorithm works on the idea that similar points are grouped together. It calculates the distance of incoming points to all existing data points. To classify the incoming point, it takes a vote based on the label of the nearest K points. The best number of neighbors to consider, or value of K, is based on the data; it will differ for each dataset. (Harrison, 2019) The image data originally had an 81% accuracy rate. After tuning the model hyperparameters of n_neighbors (value of K) and p, a parameter that specifies how the distance between points is calculated, the model had a modest 1% increase in accuracy to 82%. While that is not very good, it is higher than the average radiologist accuracy of 74% in similar image detection.

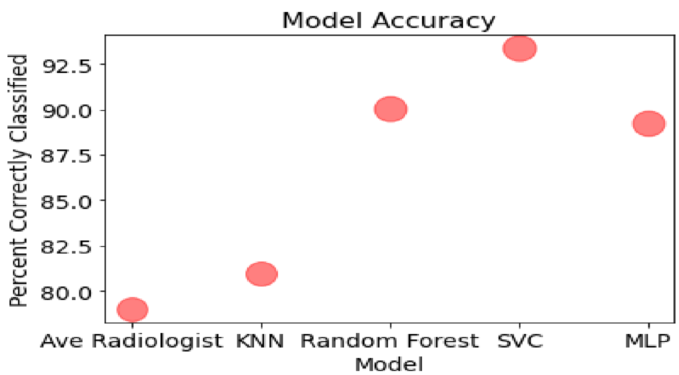
The next model applied to the images was Random Forest Classifier. Random Forest Classifier is an ensemble of decision trees. Decision trees are sensitive to the data they are trained on; that is, small differences in data create very different trees. Random Forest avoids this disadvantage by randomly selecting data from the dataset (with replacement) and randomly selecting features. The intent is to create trees with low correlation. From these collections of decision trees, each one gives a result, and the outcome with the highest number of votes gives the image its classification label. (Donges) The original result of this model was 84%. After tuning the hyperparameters of n_estimators (the number of decision trees) and max_depth (how deep the decision tree can grow) the accuracy increased to 90%. This is significantly better than our baseline radiologist result of 74%.

Next, the Support Vector Machine Classifier (SVC) was applied. SVC finds the best boundary to separates data points into different classes. The SVM creates a boundary that is maximally distant from data points in each group to generalize for new data without breaking the model. (Alto, 2019) The accuracy of the model was 93%. After tuning the hyperparameters kernel (kernel type to be used in the algorithm, the options include 'linear,' 'poly,' 'rbf,' 'sigmoid,' 'precomputed'), gamma ('auto' or 'scale', each is a different way to calculate the kernel coefficient based on the number of features), C (regularization parameter), and degree (degree of polynomial; this is ignored by all but the 'poly' kernel function). (Pedregosa, 2011) After tuning, the accuracy remained at 93%. The model defaults were found to be the best hyperparameters after tuning, therefore there was no change in the results. (Sklearn.svm.SVC)

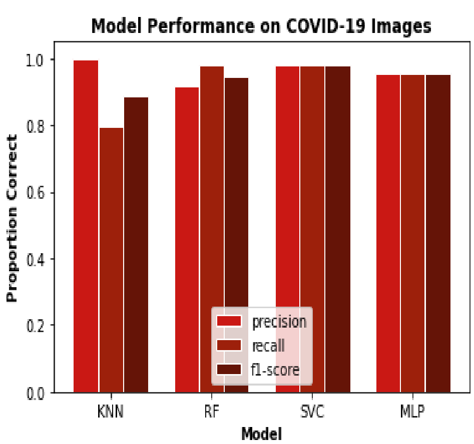
Finally, the last model trained on the data was the Neural Network Multilayer Perceptron (MLP) Classifier. An MLP is a neural network supervised machine learning algorithm. It is composed of one or more neural layers. First is the input layer, then there are hidden layers, and then predictions are made on the output layer. There are activation functions that propagate the data through network. Data is trained going backwards through the model but input data only propagates forward, as in a directed graph. (Techopedia, 2017) The original MLP model had an accuracy of 88%. The model parameters tuned were the activation function (different functions that propagate the data through the model), the solver (to determine weight optimization), alpha (L2 penalty parameter), and the learning rate (learning rate to change model weights). (Sklearn.neural_network.MLPClassifier) After tuning the hyperparameters, the accuracy went up slightly to 89%. While this is below SVM's accuracy at 93%, it is still significantly better than baseline radiologist result.

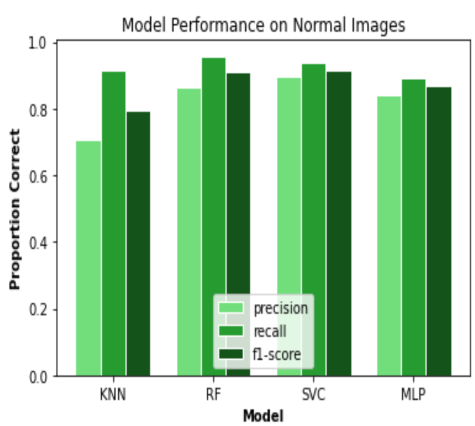
Conclusions

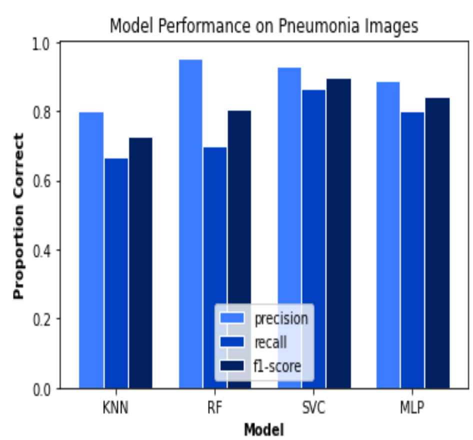
All of the models performed better than the baseline radiologist average of 74%. The accuracies achieved by the models were KNN at 82%, MLP at 89%, Random Forest at 90%, and SVC performed the best at 93%. There are certainly radiologists that have a better track record than the average radiologist from the study used as a baseline. However, the variability in radiologist diagnostic success and the human factors that make it difficult to be able to maintain high performance make this field an excellent area for machine learning augmentation.



This model accuracy diagram is a visual representation of the accuracy results. SVC performed the best at 93%, MLP and Random Forest had comparable results at 89% and 90%, and KNN was the least accurate of the machine learning models at 82%. All models performed better than the average radiologist from the study comparing pneumonia and COVID-19 lung images (74%).







The visualization above shows a comparison of how the models performed across the different types of images, COVID-19, normal, and pneumonia by the metrics of precision, recall, and F1-score. Precision is a measure of what proportion of images that were selected as a certain type of image were classified correctly. Recall considers of all images in each group in the data set, how many were correctly found to be each type. F1-score is a combination of these two measures. All of the models performed the best on the COVID-19 images. This is logical because the COVID images had the most distinguishing characteristics with the deep opacity of the lung images. The models performed less well on the normal lung images and performed the least on the pneumonia images. This may be due to the number of similarities in the images or may be due to the underrepresentation of the pneumonia images.

Further research on this topic would include adjusting the datasets to have balanced datasets in each category. The 41% fewer pneumonia images may have caused lower accuracies. Imbalanced datasets are problematic for MLP, SVC, and KNN. Random Forest is more tolerant of imbalanced datasets.

Another possible variation of this analysis would be to not perform the Histogram of Oriented Gradients (HOG) transformation on the data. SVMs trained on HOG features are known to pair well. (2014) Perhaps without this preprocessing step, the other algorithms may have had a better performance.

Additionally, Convolutional Neural Networks (CNN) would be added for study. A study was performed using CNN for pneumonia detection and achieved a remarkable 98.4% accuracy rate. (Hashmi, 2020) CNNs were specifically designed to map image data and are the standard model for image predictions. (Brownlee, 2019)

The results of this analysis show that machine learning models have tremendous promise to augment the field of radiology and improve diagnostic imaging classification.

References:

Alto, V. (2019, August 20). *Support Vector Machine from scratch*. Medium. Retrieved January 30, 2022, from https://medium.com/swlh/support-vector-machine-from-scratch-ce095a47dc5c

Bai, H. X. (2020, March 10). *Performance of Radiologists in Differentiating COVID-19 from Non-COVID-19 Viral Pneumonia at Chest CT*. Radiology. Retrieved January 29, 2022, from https://pubs.rsna.org/doi/10.1148/radiol.2020200823

Brady A. P. (2017). Error and discrepancy in radiology: inevitable or avoidable?. *Insights into imaging*, *8*(1), 171–182. https://doi.org/10.1007/s13244-016-0534-1

Brownlee, J. (2019, August 19). *When to use MLP, CNN, and RNN Neural Networks*. Machine Learning Mastery. Retrieved January 30, 2022, from https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/#:~:text=Multilayer%20Perceptrons%2C%20or%20MLPs%20for,also%20called%20the%20visible%20layer.

Canuma, P. (2020, August 28). *Image pre-processing*. Medium. Retrieved January 29, 2022, from https://prince-canuma.medium.com/image-pre-processing-c1aec0be3edf

Chan, S., & Siegel, E. L. (2019). Will machine learning end the viability of radiology as a thriving medical specialty?. *The British journal of radiology*, *92*(1094), 20180416. https://doi.org/10.1259/bjr.20180416

Collins, R. (n.d.). *Lecture 4: Smoothing - Pennsylvania State University*. Lecture 4 - Smoothing. Retrieved January 29, 2022, from http://www.cse.psu.edu/~rtc12/CSE486/lecture04.pdf

Donges, N. (n.d.). *A complete guide to the random forest algorithm*. Built In. Retrieved January 29, 2022, from https://builtin.com/data-science/random-forest-algorithm

Feature descriptor: Hog descriptor tutorial. Analytics Vidhya. (2020, May 10). Retrieved January 29, 2022, from https://www.analyticsvidhya.com/blog/2019/09/feature-engineering-images-introduction-hog-feature-descriptor/

Harrison, O. (2019, July 14). *Machine learning basics with the K-nearest neighbors algorithm*. Medium. Retrieved January 29, 2022, from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Hashmi, M. F., Katiyar, S., Keskar, A. G., Bokde, N. D., & Geem, Z. W. (2020). Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning. *Diagnostics (Basel, Switzerland)*, *10*(6), 417. https://doi.org/10.3390/diagnostics10060417

Image analysis in python. Introduction to three-dimensional image processing - Image analysis in Python. (n.d.). Retrieved January 29, 2022, from https://scikit-image.org/skimage-tutorials/lectures/three_dimensional_image_processing.html

Queensland University of Technology, Australia. (2014, June 10). *Why do linear svms trained on hog features perform ... - arxiv*. e-Print archive. Retrieved January 30, 2022, from https://arxiv.org/pdf/1406.2419.pdf

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

scikit learn. (n.d.). *Sklearn.svm.SVC*. sklearn. Retrieved January 30, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Shams, Mahmoud; Elzeki, Omar; Abd Elfattah, Mohamed (2020), "Chest X-ray images with three classes: COVID-19, Normal, and Pneumonia", Mendeley Data, V1, doi: 10.17632/fvk7h5dg2p.1

Sklearn.neural_network.MLPClassifier. scikit. (n.d.). Retrieved January 30, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html?highlight=mlp#sklearn.neural_network.MLPClassifier

Techopedia. (2017, March 30). *What is a multilayer Perceptron (MLP)? - definition from Techopedia*. Techopedia.com. Retrieved January 30, 2022, from https://www.techopedia.com/definition/20879/multilayer-perceptron-mlp#:~:text=A%20multilayer%20perceptron%20(MLP)%20is,the%20input%20and%20output%20layers.&text=MLP%20is%20a%20deep%20learning%20method.