Project Title: **Interpretable Machine Learning And Data Augmentation: A Viable Application for Medical Image Analysis in a Clinical Setting**

Team Member Names: Joshua Kraus
Siddharth Perini
Jennifer Sentiff

Introduction

As one of the quickest-developing fields in machine learning and artificial intelligence, deep learning can be utilized in numerous disciplines, namely medicine, to improve process efficiency (*Suganyadevi et al., 2022*). Through tasks such as cancer screening and infection monitoring, deep learning can be utilized to aid doctors in their personalized care and treatment of patients while reducing the time and resources required to do so (*Zhang et al., 2016*). With a dramatic increase in requests for medical imaging services and a shortage of radiologists, a strain has been put on the healthcare system to address patients' needs in a timely manner (*Puttagunta & Ravi, 2021*). The use of machine learning for medical image analysis is one way to address this problem, as it could lessen the manual time and resources required to analyze images (*Puttagunta & Ravi, 2021*). Notably, however, the lack of large amounts of properly annotated data has limited the advancement of deep learning for medical image analysis (*Minaee et al., 2020*). Additionally, another major barrier to implementing deep learning models in this context is the concern surrounding the use of uninterpretable black-box models in healthcare (*Suganyadevi et al., 2022*). As a solution to this, interpretable supervised learning models with easily obtainable tracking and measurements are possibly a feasible way to implement machine learning in a clinical setting (*Price, 2018, Suganyadevi et al., 2022*). Moreover, data augmentation techniques that synthetically create more training data for a model could potentially improve the performance of any machine learning models used where a lack of training data may exist (*Chlap et al. 2021, Suganyadevi et al., 2022*). The combination of an interpretable machine learning model and training data augmentation could provide an approach to medical image analysis that addresses issues of validation for healthcare professionals and issues of regulation by government officials. If such an approach is reasonably accurate enough, its use in a clinical setting could provide medical experts with a tool that informs them during their medical image analysis process to reduce the time and resources required to do so, while possibly decreasing errors and variability in diagnoses (*Marias, 2021; Puttagunta & Ravi, 2021*).

Review of Related Literature

Artificial Intelligence in Medical Image Analysis
Deep learning is a popular field in machine learning and artificial intelligence, with possible use in medicine to streamline certain processes (*Suganyadevi et al., 2022*). In healthcare, machine learning models can be implemented for clinical information analysis and image examination to

positively affect patients' lives with little effort or time required (*LeCun et al., 205*). Notably, the process of medical image analysis is a costly one and may involve radiologists consulting colleagues or previous cases before diagnosis (*Hassan et al., 2020*). With ongoing improvements in deep learning techniques, however, artificial intelligence models are able to identify and classify patterns within clinical images in a much shorter time (*Chollet et al., 2015*). Through tasks such as cancer screening and infection monitoring, deep learning can be utilized to aid doctors in their personalized care and treatment of patients while reducing the time and resources required to do so (*Zhang et al., 2016*).

In addition to the possible benefits of artificial intelligence for medical image analysis, certain deep learning architectures for image processing have been found to be highly accurate for classification and detection (*Ahuja, 2020; Bengio, 2012; Singh, 2020*). Most notably, the ResNet architecture is a common choice for a convolutional neural network which produces results with high accuracy for image analysis (*Bengio, 2012*). In the context of COVID-19 classification and detection, variations of the ResNet architecture and other convolutional neural networks have produced accuracy levels as high as 99.40% and 93.29%, respectively (*Ahuja, 2020; Singh, 2020*).

Benefits of Machine Learning in a Clinical Setting

In the current state of the healthcare system, there is a larger demand for medical image services than there are radiologists to analyze them (*Puttagunta & Ravi, 2021*). With a dramatic increase in requests for medical imaging services and a shortage of radiologists, a strain has been put on the healthcare system to address patients' needs in a timely manner (*Puttagunta & Ravi, 2021*). This strain is compounded by the fact that analyzing medical images is a lengthy, time-consuming process for radiologists to undergo, which only furthers this issue (*Puttagunta & Ravi, 2021*). The use of machine learning for medical image analysis is one way to address this problem, as their use could lessen the manual time and resources required to analyze images (*Puttagunta & Ravi, 2021*). While not a direct replacement for human analysis, the use of artificial intelligence in this manner could be implemented as an initial screening for radiologists to streamline their analysis process.

In addition to the cost-benefit of utilizing machine learning for medical image analysis, given a model is accurate enough, it could help reduce the number of missed detections and settle discrepancies among radiologists as well (*Marias, 2021*). One of the main benefits of machine learning is that its image analysis capabilities exceed the limits of human vision, which could potentially provide highly accurate results for medical image analysis comparable to manual examination by radiologists (*Marias, 2021*). Given the possible high degree of accuracy of machine learning models in this context, their results and predictions could also be used to cope with any inter-observer variability in diagnoses of medical images (*Marias, 2021*). While still not a replacement for radiologists, given a machine learning model is accurate enough for medical image analysis, this shows how they could provide insight to mitigate human error and variability as well as decrease the cost and resources required to do so.

Challenges with Deep Learning for Medical Image Analysis

While deep learning models have proven benefits and high accuracy levels for medical image analysis, there are challenges with applying artificial intelligence in healthcare that serve as significant barriers to implementation (*Civit-Masot et al., 2020; Minaee et al., 2020*). Notably, the lack of large amounts of properly annotated data has limited the advancement of deep learning for medical image analysis (*Minaee et al., 2020*). Especially when compared to general computer vision data sets, medical imaging data is far too limited for the purposes of applying artificial intelligence (*Apostolopoulos & Mpesiana, 2020*). This has led to the overfitting of training data sets, an issue that limits the capabilities of deep learning in this context (*Suganyadevi et al., 2022*). There has been an increasing trend in the medical community to

make large annotated data sets available, but it is still unclear how well artificial intelligence can perform given the current data constraints (*Panwar et al., 2020*).

In deep learning, artificial neural networks form the basis for most architectures and have some similarities to biological neural networks (*Suganyadevi et al., 2022*). Neural networks are composed of multiple layers that transform the input into output by processing and learning new, nonlinear information (*Apostolopoulos & Mpesiana, 2020*). With each layer utilized, the data is transformed into a higher, more abstract level meaning deeper in the network more complex information is learned (*Apostolopoulos & Mpesiana, 2020*). Deep learning refers to deeper networks than simpler machine learning, which can be utilized to analyze highly complex and changing data (*Apostolopoulos & Mpesiana, 2020*). The drawback of utilizing deep learning architectures, however, is their "black-box" approach where due to the highly complex network and neurons and layers, it becomes impossible to fully explain the process a deep learning model undergoes to turn input into output or how it changes in response to new data (*Suganyadevi et al., 2022*).

In addition to concerns about the quantity of data available, another major barrier to implementing deep learning models for medical image analysis is the concern surrounding the use of black-box models in healthcare (*Suganyadevi et al., 2022*). In deep learning, artificial neural networks form the basis for most architectures and have some similarities to biological neural networks (*Suganyadevi et al., 2022*). Neural networks are composed of multiple layers that transform the input into output by processing and learning new, nonlinear information (*Apostolopoulos & Mpesiana, 2020*). With each layer utilized, the data is transformed into a higher, more abstract level meaning deeper in the network more complex information is learned (*Apostolopoulos & Mpesiana, 2020*). Deep learning refers to deeper networks than simpler machine learning, which can be utilized to analyze highly complex and changing data (*Apostolopoulos & Mpesiana, 2020*). The drawback of utilizing deep learning architectures, however, is their "black-box" approach where due to the highly complex network and neurons and layers, it becomes impossible to fully explain the process a deep learning model undergoes to turn input into output or how it changes in response to new data (*Suganyadevi et al., 2022*). Given this understanding, black-box models like deep learning can be viewed as inherently opaque in their method and fluid in their interpretation, raising questions regarding the validation and regulation of these models (*Price, 2018*). Artificial intelligence techniques often provide a prediction or recommendation but do not explain how those results were exactly derived for justification (*Price, 2018*). This is contrasted by simpler analytical models which may be more interpretable and raises the question of whether machine learning models in healthcare should be pushed towards interpretable models aimed at increasing understanding or black-box algorithms aimed at practicality (*Price, 2018*). To compound this problem, black-box algorithms often change and adapt in response to new, varied data, an uncommon approach in medicine (*Price, 2018*). While frequent updates are common in software and analytics, this is relatively rare in healthcare where, for example, certain verified drugs have been used for centuries (*Price, 2018*).

Both the issue of opaqueness and plasticity provide challenges when pitching the implementation of deep learning models in medicine to healthcare professionals and government officials. Medical professionals may question how black-box models can be ensured for accuracy and use, or, if a model can accurately measure a clinical quantity of interest that can usefully guide clinical care (*Price, 2018*). Whereas diagnostic tests would typically be used to determine this, it cannot be precisely determined what deep learning models measure and track due to their opaqueness (*Price, 2018*). Moreover, the plasticity of artificial intelligence compounds this issue since as models change in response to new data, it becomes impossible to rely on understanding to provide confidence in the efficacy of a model. Furthermore, the use of black-box models has legal ramifications that act as a deterrent for healthcare professionals and ultimately a barrier

for government officials (*Suganyadevi et al., 2022*). Medical algorithms are under the jurisdiction of the Food and Drug Administration, the FDA, which requires the demonstration of the safety and efficacy of static potential products and prohibits those which do not allow providers to review the justification for their results (*Price, 2018*). Given that black-box models cannot be interpreted and evolve to new data, these specifications by the FDA broadly ban the use of black-box models for medical image analysis, at least in their current form. Given that the solutions to these challenges are not readily apparent, for the foreseeable future deep learning models are unfeasible for the purposes of medical image analysis.

Possible Solutions to Challenges in Medical Image Analysis

Given that black-box algorithms such as deep learning models are all but infeasible for the purposes of medical image analysis, the use of interpretable, mechanistically modeled algorithms that are aimed at increasing understanding serves as a possible alternative for image processing. While not a replacement for radiologists, such models could be feasibly implemented, unlike deep learning, to inform medical experts about the images they are examining and streamline their analysis process (*Puttagunta & Ravi, 2021*). Additionally, while larger, annotated data sets may become available in the coming years, another viable alternative to combat the issue of limited data to sufficiently train models is the use of data augmentation to synthetically increase the size of data sets for model fitting (*Minaee et al., 2020, Suganyadevi et al., 2022*). The use of data augmentation could improve the performance of machine learning models on a separate validated data set by providing additional data to train the model for less cost than collecting a comparable amount of new data (*Chlap et al., 2021*). In this manner, interpretable models and data augmentation techniques serve as possible solutions to the challenges of implementing machine learning models in a clinical setting, both of which require further discussion.

Related most closely to the task of deep learning models, interpretable supervised learning models exist, which could be feasibly implemented for medical image analysis (*Suganyadevi et al., 2022*). Defined as a model where the representation between input examples and a target variable is learned, supervised learning problems consist of classifying a class or numeric label (*Suganyadevi et al., 2022*). Neural networks for deep learning are an example of this, as are decision trees, support vector machines, logistic and linear regression as well as the k-nearest neighbor algorithm (*Suganyadevi et al., 2022*). For the purposes of medical image analysis, supervised learning models could be fed a vector of images as input data with specific disease classifications as the target variable to train a model which can predict the classification of a new image. In each of these models, other than neural networks, what the supervised learning model tracks and measures in order to make predictions can be easily obtained and examined, making them possibly feasible for implementation in a clinical setting (*Price, 2018, Suganyadevi et al., 2022*). While the exact performance of each of these models for medical image analysis has not been tested or compared to the high degree of accuracy of deep learning models, it is possible that an interpretable supervised learning model exists which is also accurate for this task.

Various data augmentation techniques exist as well to generate synthetic data, some of which are cost-efficient, elementary techniques (*Chlap et al. 2021, Suganyadevi et al., 2022*). Such techniques are centered around transforming an image in such a way to map the points to a different position or manipulate intensity values to create a synthetic image (*Chlap et al. 2021*). Examples of this include geometric transformations to the position of the image points, cropping or occluding the image, altering the intensity of the image points, injecting noise into the image, sharpening or blurring the image through filtering, and performing principal component analysis on the image (*Chlap et al. 2021, Suganyadevi et al., 2022*). Through any of these processes, it is possible to significantly increase the size of a training dataset, which could potentially improve the performance of machine learning models. Moreover, by augmenting the model training data, it is possible to introduce more variance in the image data used for fitting the models, and

possibly reduce overfitting (*Suganyadevi et al., 2022*). The tradeoff between choosing between data augmentation techniques must be noted, however, as the basic techniques previously discussed are easier to implement than more complex approaches such as generating data using deep learning through a GAN, but these elementary approaches also may not yield as positive of an effect as more intricate techniques (*Suganyadevi et al., 2022*).

Problem Statement

Need for the Study

Given that the use of machine learning models for medical image analysis faces issues of interpretability and training limitations, more research is required to offer potential twofold solutions to these issues. The use of interpretable supervised models provides a viable alternative to black-box models by being able to examine what the algorithm measures and tracks with a static model, making it possible to run diagnostic tests and clinical trials on its implementation (*Price, 2018*). Moreover, elementary data augmentation techniques have shown the possible ability to improve the performance of machine learning models by providing additional, varied data to train the model which is then validated on a separate data set (*Chlap et al., 2021*). While the use of interpretable supervised models for medical image analysis and data augmentation techniques have both been examined separately in research, the synthesis of these ideas to create a possible accurate, and feasible model for clinical implementation remains relatively untested. Therefore, more research is required to determine the useability and accuracy of such a strategy for medical image analysis. If successful, the use of an interpretable model and data augmentation could provide an approach to medical image analysis that addresses issues of validation for healthcare professionals and issues of regulation by government officials. Moreover, the possible time and resource benefits implementing a machine learning model could have for radiologists requires that an approach be developed which could be implemented in a clinical setting. If such an approach is accurate enough, its use in healthcare could provide medical experts with a tool that screens images and informs radiologists during their examination to reduce the time and resources required to do so while possibly decreasing errors and variability in diagnoses (*Marias, 2021; Puttagunta & Ravi, 2021*).

Purpose of the Study

The purpose of this study is to propose a medical image analysis framework to address the concern of black-box models in healthcare by implementing interpretable supervised learning models, and approach the issue of limited training data by utilizing data augmentation techniques. Specific research questions include:

- Which machine learning model performs the best? How accurate is it, and how does this compare to deep learning models for medical image analysis?
- What does the best-performing model track and measure, and how can it be interpreted?
- Which data augmentation technique performs the best? What is its effect on the best-performing model?

Data Source

The NIH chest X-ray dataset contains 112,120 frontal-view X-ray images, pertaining to 30,805 unique patients, extracted from the clinical PACS database at the National Institute of Health (NIH) Clinical Center and representing about 60% of frontal chest X-rays collected at that location. This dataset is considered by NIH as the most representative of real population distributions, compared to any previous chest X-ray data sets. The full data set is split amongst twelve tar.gz compressed files, each containing 1-4 GB of PNG image files in 1024x1024 resolution. Metadata for all images are provided in a CSV file (*Data_Entry_2017.csv*) that includes an image index, finding labels, follow-up #, patient ID, patient age, patient gender,

view position, original image size, and original image pixel spacing. Each image is associated with one of fourteen disease labels including Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule, Mass, and Hernia. These disease labels were text-mined from patient radiology reports and are expected to have an accuracy of > 90%. The full dataset can be found here: https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345

Methodology

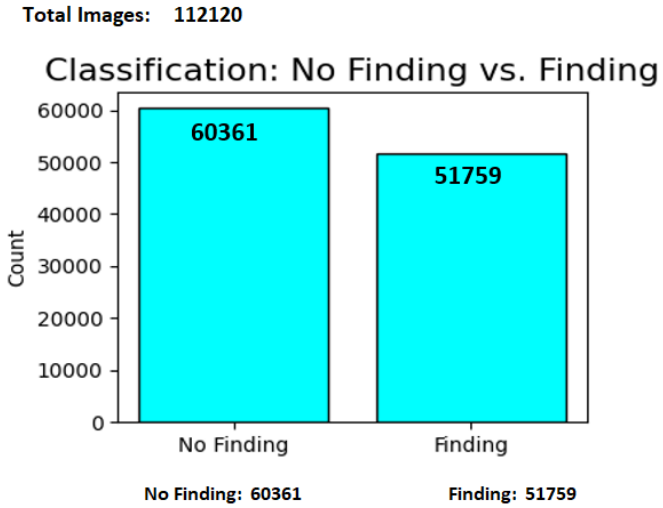
Preparing Data

The NIH X-ray data were downloaded using Google Colab, uncompressed using Python's tarfile module, and stored on Google Drive as Python NumPy array files (.npy). The image data array file contains 112,120 unique chest X-ray images, such that each column of the data array corresponds to a 64x64 pixel image vector $\in \Re^{4096}$. The data labels array file is one-dimensional and contains the disease classifications corresponding to the 112,120 unique chest X-ray (CXR) images. This array data was processed from the clinical findings in the *Data_Entry_2017.csv* file and mapped to their respective images in the CXR data set. Creating data and label arrays for this image data set proved to be simpler than storing all 45 GB of data on local machines. Similarly, to avoid the memory requirements when processing data, using this approach addressed size constraints when reading in the X-ray data. Moreover, loading the data matrix in this format reduced computational cost, time, and reduced storage costs (>50% reduction in size compared to .csv & .txt files), making it a reasonable choice for this context.

Exploratory Data Analysis

Upon initial examination of the data, we found that of the total 112,120 images, 60,361 were classified with no finding and 51,759 were classified as having a pathology finding, which can be seen in Figure 1 below.

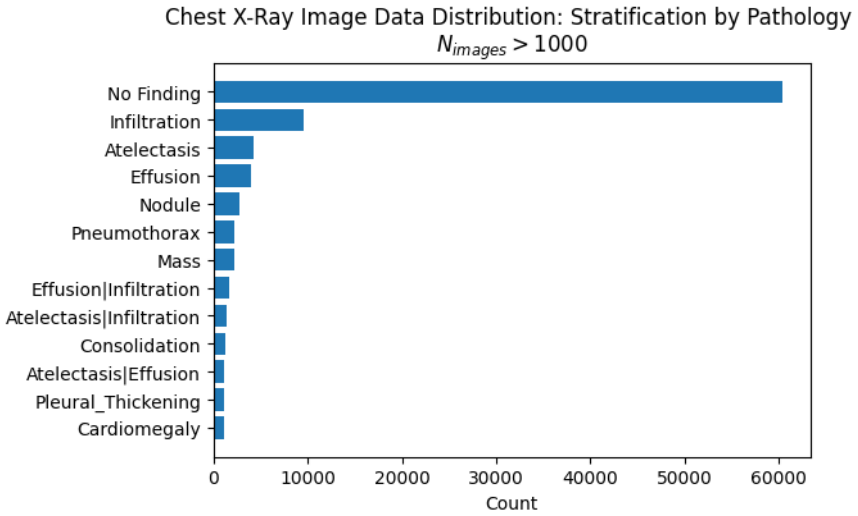
Figure 1: Chest X-ray image data set classification - No Finding vs. Finding



To further evaluate the distribution of the data, pathologies with less than 1000 images were excluded from the data set, and the individual frequencies of each pathology finding were examined in Figure 2 below.

Figure 2: Chest X-ray image data set classification - Pathologies (N_{images} > 1000)

Image Classification N_{images} >1000		Count	% of Total
No Finding		60361	65.07%
Finding	Infiltration	9547	10.29%
	Atelectasis	4215	4.54%
	Effusion	3955	4.26%
	Nodule	2705	2.92%
	Pneumothorax	2194	2.37%
	Mass	2139	2.31%
	Effusion|Infiltration	1603	1.73%
	Atelectasis|Infiltration	1350	1.46%
	Consolidation	1310	1.41%
	Atelectasis|Effusion	1165	1.26%
	Pleural_Thickening	1126	1.21%
	Cardiomegaly	1093	1.18%
	Total	92763	



To begin, initial testing took place to determine the feasibility of modeling various pathologies for binary classification. After weighing the performance of classification models and the size of the data sets for various pathologies, it was determined that machine learning models for binary classification of the CXR images for "No Finding" or finding = "Effusion" would be the most feasible. When this pathology was used for binary classification all models performed better compared to other pathologies, while still being the third most frequent pathology classification found in the data set.

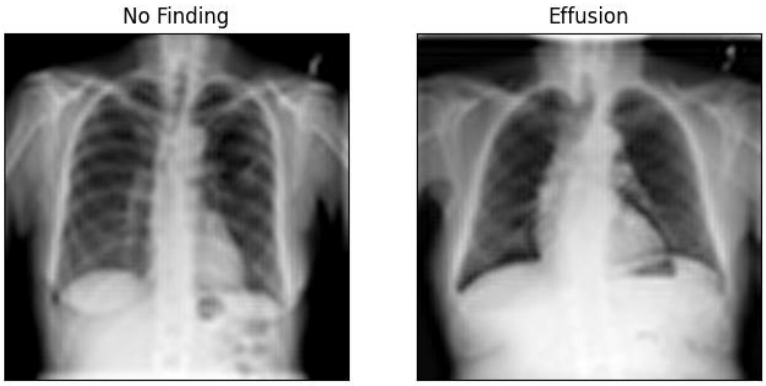
After subsetting the data set to only include classifications of "No Finding" or "Effusion", issues of imbalanced data became apparent. The table in Figure 3 demonstrates this imbalance as 6.15% of the images are those labeled as class Effusion and the remaining 93.85% of the images are of the class "No Finding".

Figure 3: Data Distribution - No Finding and Finding (Effusion)

Image Classification N_{images} >1000		Count	% of Total
No Finding		60361	93.85%
Finding	Effusion	3955	6.15%
	Total	64316	

Additionally, shown in Figure 4 are images of chest radiographs from the CXR data set for each class used in this analysis.

Figure 4: Chest X-ray Images - No Finding and Effusion



Data Set Balancing

To examine the possible issues of training and testing classifiers on imbalanced data, both the imbalanced data set containing all "No Finding" or "Effusion" images and a balanced data set achieved through downsampling the "No Finding" images were tested.

For the imbalanced data set, a total of 64,316 images were used with 60,361 labeled as "No Finding" and 3,955 labeled as "Effusion". This data set was then split into training and testing sets, with 80% of the data used for training and the remaining 20% for testing. Each of the classifiers were then trained on the 51,452 images in the training set, and then predictions were made on the 12,864 images in the testing set to examine each model's performance. For this testing set, 12,099 images were classified as "No Finding" while 765 images were classified as "Effusion". It should be noted that a possible issue with training classifiers on this data set is decreased model performance due to an underlying assumption of balanced class distribution and equal misclassification costs (*Sun et al., 2009*). Given this possible concern, the performance of models fit on the imbalanced data set were examined to determine the effect of unequal class distributions.

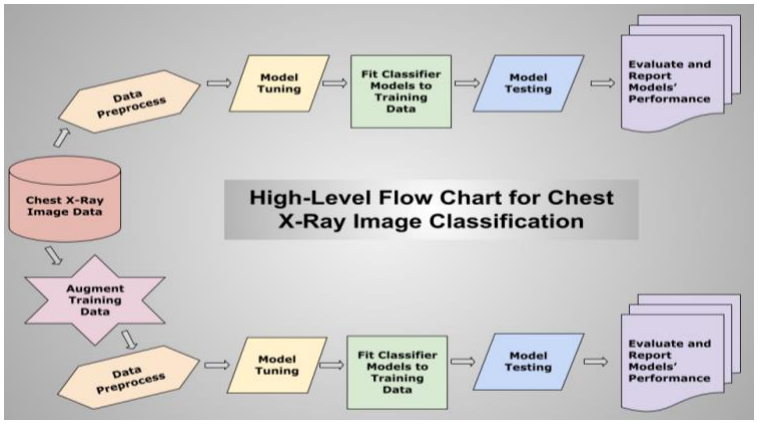
For the balanced data set, the "No Finding" images were downsampled to match the number of "Effusion" images in the data set. This resulted in a data set totaling 7,910 images, with 3,955 coming from each classification. This data set was then split into training and testing sets where the models were trained on 6,328 images (80% of the data), and predictions were made on the remaining 20% 1,582 images, to assess each classifier's performance. For these splits, each set was stratified to ensure that the frequency of each classification remained balanced in both sets resulting in a testing set with both 791 "No Finding" and "Effusion" images. This data set attempts to correct the issue of imbalanced class distributions but at the cost of using far less training data, therefore the performance of models fit on this data set were examined to determine the effect of using less training data with equal class distributions.

Classification Models

Several supervised classification models were implemented such as (Artificial) Neural Networks (ANN), K-Nearest Neighbors (KNN), Kernel-Support Vector Machine (K-SVM), Naïve Bayes Classifier, Logistic Regression Classifier, and Gradient Boosting, along with unsupervised clustering models like K-Means, and unsupervised dimensionality reduction methods like Principal Component Analysis (PCA).

For each model, various standard or widely used ML preprocessing techniques were performed on the chest X-ray image data. These include flattening image arrays (vectorizing the image data), shuffling rows, downsampling the image resolution, converting to grayscale, and binarizing the y-response. For the ANN model, initialization steps included selecting the initial number of layers, activation function, and L2 regularization term, while tuning model parameters in increments. When implementing the KNN model different K-values, distance formulas, and weight functions were tested. Multiple K-SVMs with different kernel types, C-values, and shrinking heuristics were also trained and tuned. For Logistic Regression, various penalty terms, regularization strengths, and optimization algorithms were examined. The Gaussian Naïve Bayes classifier with default prior probabilities was implemented with tuned smoothing parameters. For the Gradient Boosting model, the learning rate and the number of boosting stages were tested to evaluate acceptable performance. Multiple K-Means models with varying initialization methods and solver algorithms were also examined to find the best approach. Lastly, for our unsupervised PCA model, the goal was to use dimensionality reduction techniques to reduce computational cost while hopefully retaining comparable model performance. Multiple models were fit with a specified number of principal components to achieve this task. Figure 5 demonstrates a high-level approach to the methods and techniques implemented in the work reported here.

Figure 5: ML Approach for Classification using CXR Images.



Model Tuning

Models were each tuned using a randomized search of ranges of various hyperparameters. In this approach, all specified parameters were searched simultaneously by selecting random combinations of hyperparameters, which drastically lowers the run time of a randomized search compared to testing all possible combinations in a full grid search. While a randomized search is not guaranteed to provide the best possible combination of hyperparameters, it often provides comparable performance to a full grid search with slight discrepancies in performance most likely being due to a noise effect that would not apply to a testing data set. Given this understanding, a randomized search is a feasible alternative to a full grid search that will provide a reasonably tuned model in a fraction of the time. The full ranges of hyperparameters tested for each classifier can be seen in Figure 6. For full documentation of each classifier model, please visit scikit-learn's website found here: https://scikit-learn.org/stable/user_guide.html.

Figure 6: Hyperparameter Tuning for scikit-learn Classifier Models

ANN		**KNN**		**SVM**		**Logistic Regression**	
Parameter	**Range**	**Parameter**	**Range**	**Parameter**	**Range**	**Parameter**	**Range**
hidden_layer_size	(10,), (50,), (100,), (200,)	n_neighbors	50-200	C	0.01, 0.1, 1, 10	penalty	l1, l2, elasticnet
activation	tanh, relu	weights	uniform, distance	kernel	poly, rbf	C	0.01, 0.1, 1
alpha	0.001, 0.01, 0.1	p	1, 2	shrinking	True, False	solver	sag, saga

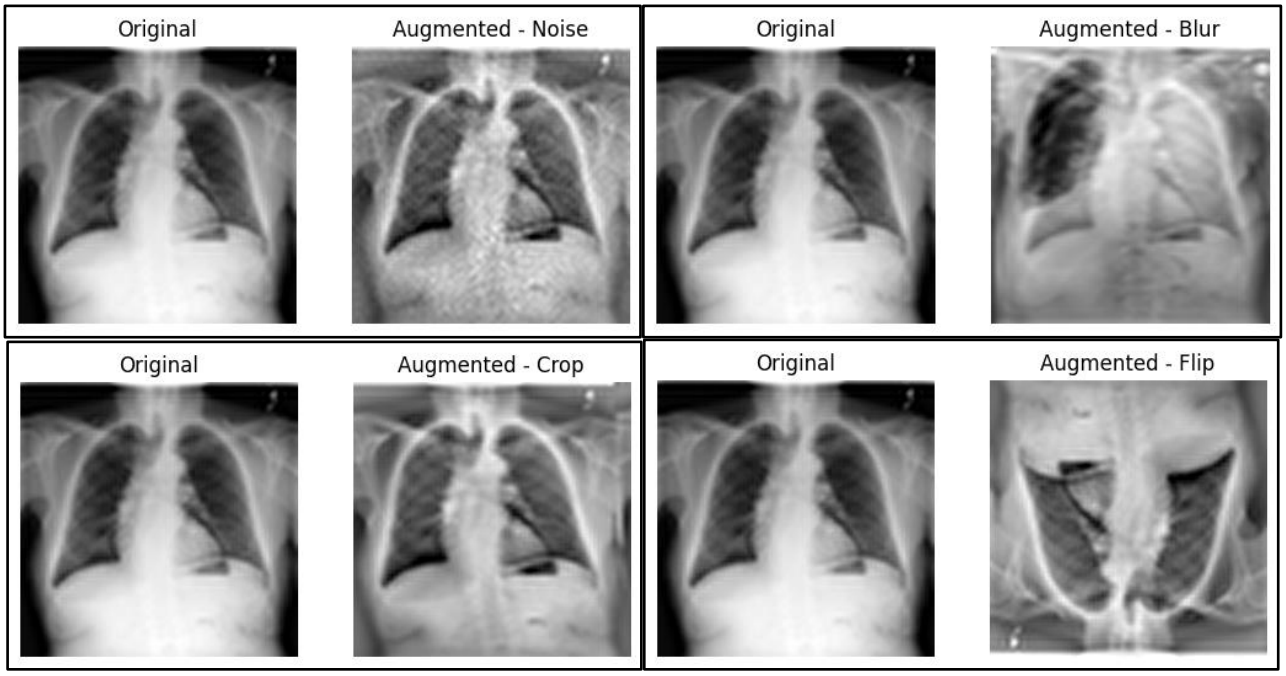
Naïve Bayes		**Gradient Boosting**		**K-Means**	
Parameter	**Range**	**Parameter**	**Range**	**Parameter**	**Range**
var_smoothing	1e-9, 1e-7, 1e-5, 1e-3, 1e-1, 1e0	n_estimators	100, 200, 300, 400, 500	algorithm	full, elkan
		learning_rate	0.01, 0.1, 1	n_init	10, 20, 30, 40

Data Augmentation

To attempt to improve the performance of various classifiers, Python's ImgAug library was used to perform various image transformations on the original dataset to generate augmented training data sets. The main focus of this augmentation process was to increase the size of the training data set which had a notable imbalance between class distributions. By augmenting the "Effusion" images, the total number of images of this classification could be increased in the training data set and balanced with the "No Finding" images to create a larger training data set. Additionally, by augmenting the existing images in this data set, the training data set would then account for more variance in chest radiographs with the pathology "Effusion". Through this process, the goal of augmenting the "Effusion" images in the training set was to improve model performance by providing a larger data set for model training which better accounted for variance across these types of thoracic X-ray images.

Four main augmentation techniques were implemented which resulted in four new data sets. The augmented data sets included noise injection, blurring, cropping, and geometric transformations, such as horizontal and vertical flipping on the training data. These four augmented data sets were then used to train each of the machine learning models. Afterward, the performance of each model using each augmented training set was examined on the testing data set to determine the effect each augmentation technique had on model performance. An example of the image augmentation methods implemented can be seen in Figure 7.

Figure 7: Sample Images - Data Augmentation Techniques



Classifier Performance Evaluation

Different classifiers were compared with and without augmented training data, and each model's performance metrics for the detection and/or presence of lung disease using the NIH Chest X-ray image dataset were reported. Augmented training data generated via geometric transformations, image cropping, image blurring, and image noise injection were used to train our machine learning classifiers; and each of these augmented training sets were compared to the models fit on the non-augmented data for baseline comparison.

The precision, recall/sensitivity, accuracy, F1-score, and Area Under Curve scores (AUC) for each of the classifiers were reported with and without data sets using data augmentation, as well as for combined methods such as using balanced and imbalanced data for training. Furthermore, PCA was performed on the training data for dimensionality reduction to n=10 dimensions. The explained variance ratio for the first 10 principal components was 62.9%, however, PCA was eventually not used on the image data as the initial testing yielded poor model performance for the classification tasks. Additionally, the Receiver Operating Characteristic curves (ROC) were plotted with the true positive rate (TPR) on the horizontal axis and the false positive rate (FPR) on the vertical axis. The definitions of each of the performance metrics which will be included in the performance summary can be seen below:

$$Precision = \frac{TP}{TP + FP} \qquad Sensitivity = \frac{TP}{TP + FN} \qquad Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad F1\ Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \qquad TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative, TPR = True Positive Rate, FPR = False Positive Rate

Through this process, the effect of various data augmentation techniques on model performance, as well as the overall performance of each of the classifiers for detecting the pathology of effusion in chest X-ray images can be examined. The final discussion attempted to find a reasonable method of comparison for the findings of this work to previously reported deep learning models trained on the NIH chest X-ray data set. Metrics for comparing results across previous work on this data, also referred to as the *ChestX-ray14* data set, involve computing the AUC score for each disease class (*Chen, K., et al., 2022).* The summary tables in the previously reported work (*Chen, K., et al., 2022)* provide a comprehensive AUC score comparison for various studies using DL and serve as a suitable benchmark for the results shown in our work.

Results

Balanced Non-Augmented Data

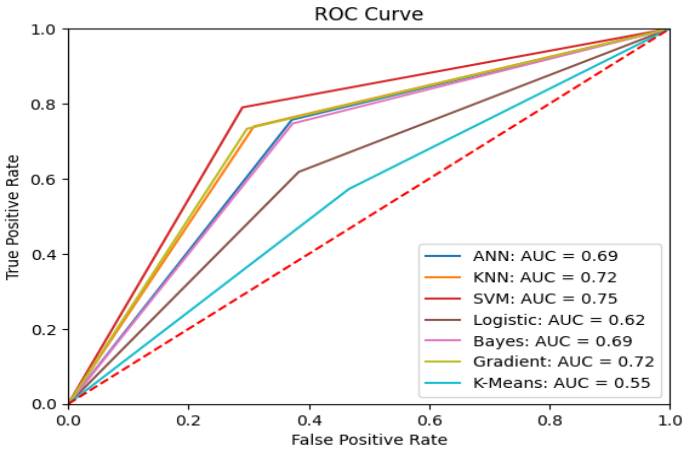
The balanced non-augmented data set consisted of a total of 7,910 images, 80% of which were used for training and the remaining 20% for testing. The seven classification models, ANN, KNN, kernel-SVM, Logistic Regression, Naïve Bayes, Gradient Boosting, and K-Means were trained on 6,328 images and then tested on 1,582 images. Of the seven models tested, kernel-SVM had the best performance with an AUC score of 0.75 and model test accuracy of 75.0%. Additionally, the ANN, KNN, Naïve Bayes, and Gradient Boosting models all achieved comparable precision scores, but only the KNN and Gradient Boosting models also had reasonable sensitivity scores. Both of these models performed relatively well when considering their F1-Score and accuracy, but neither the KNN or Gradient Boosting model performed comparably to the kernel-SVM model. Figure 8 summarizes each model's performance results and their respective ROC curves.

Figure 8: Performance Summary & ROC Curves of Classifier Models - Balanced Non-Augmented Data

Summary of Results for Balanced Non-Augmented Data

Model Performance Results - Balanced Non-Augmented Training Data

DISEASE CLASS = EFFUSION			Precision	Sensitivity/Recall	F1-Score	Accuracy	AUC
Supervised Learning	Classification	ANN	0.75727	0.670773	0.7114	0.69279	0.69
		KNN	0.73957	0.705669	0.72222	0.71555	0.72
		kernel-SVM	0.79014	0.73185	0.75988	0.75032	0.75
		Logistic Regression	0.61821	0.617424	0.61781	0.61757	0.62
		Naïve Bayes	0.74716	0.667043	0.70483	0.68711	0.69
		Gradient Boosting	0.73325	0.711656	0.72229	0.71808	0.72
Unsupervised Learning	Clustering (Classification)	K-means	0.57269	0.551095	0.56169	0.5531	0.55
	Dimensionality Reduction	PCA	*No dimensionality reduction was performed on the data used to fit and test these models.*				



Imbalanced Non-Augmented Data

The imbalanced non-augmented data set consisted of a total of 64,316 images, 80% of which were used for training and the remaining 20% for testing. The seven classification models were trained on 51,452 images and then tested on 12,864 images. The imbalanced test data set had 765 "Effusion" images and 12,099 "No Finding" images, that is, 5.95% of the test set was of the class "Effusion". Of the seven models tested, the Naïve Bayes classifier had the best performance with an AUC score of 0.68 and model test accuracy of 63.8%. This model notably still had a lower precision, accuracy, and AUC score than the Naïve Bayes classifier fit on the balanced non-augmented data set. The KNN, kernel-SVM, and Gradient Boosting models incorrectly classified every "Effusion" image in the test set as "No Finding", and correctly classified every "No Finding" image, which explains how each model's accuracy was 94% on the test data. This demonstrates the need for a balanced data set to train and tune models for binary classification which will perform reasonably. Also, this imbalance of data shows that the

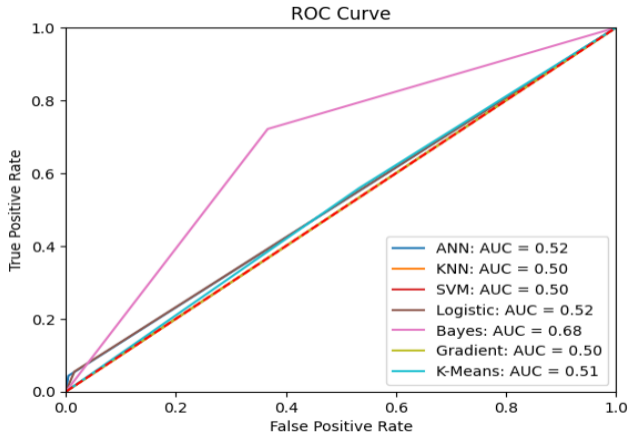
models' test accuracies are misleading, and proves why other performance metrics should be reviewed when evaluating overall performance. Figure 9 summarizes each model's performance results and their respective ROC curves.

Figure 9: Performance Summary & ROC Curves for Classifier Models - Imbalanced Non-Augmented Data

Summary of Results for Imbalanced Non-Augmented Data

Model Performance Results - Imbalanced Non-Augmented Training Data

		DISEASE CLASS = EFFUSION	Precision	Sensitivity/Recall	F1-Score	Accuracy	AUC
Supervised Learning	Classification	ANN	0.043137	0.383721	0.077556	0.938977	0.52
		KNN	0	0	0	0.940532	0.50
		kernel-SVM	0	0	0	0.940532	0.50
		Logistic Regression	0.054902	0.184211	0.084592	0.929338	0.52
		Naïve Bayes	0.721569	0.110621	0.191833	0.638448	0.68
		Gradient Boosting	0	0	0	0.940532	0.50
Unsupervised Learning	Clustering (Classification)	K-means	0.559477	0.062191	0.111939	0.472093	0.51
	Dimensionality Reduction	PCA	*PCA was only used to initally examine the explained variance ratio in the projected dimensions (the first 10 principal components). The amount of variance explained by the first 10 PCs = 62.9%. It was decided to train and fit the models on the full imbalanced data set and not use the reduced dimentional data via PCA for model fitting, as the PCA models performed poorly.*				



Balanced Augmented Data Set - Additive Noise

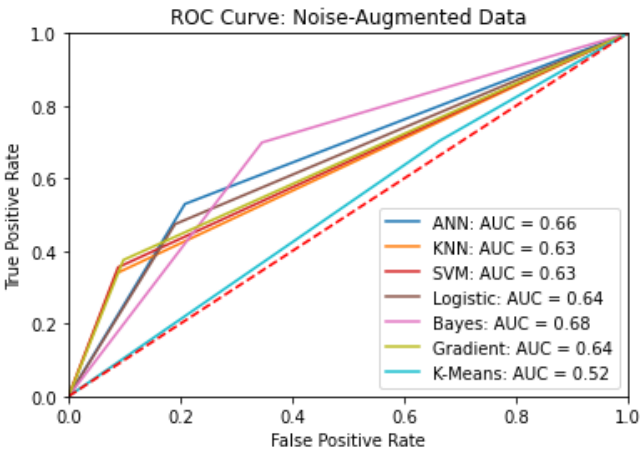
The balanced augmented data with additive noise consisted of a total of 15,820 images, 80% of which were used for training, and the remaining 20% for testing. The seven classification models were trained on 12,656 images and then tested on 3,164 images. Of the seven models tested, the Naïve Bayes classifier had the best performance with an AUC score of 0.68 and model test accuracy of 66.9%. Notably, this model had a lower precision, accuracy, and AUC score than the Naïve Bayes classifier fit on the balanced non-augmented data set. In general, when considering F-1 and AUC scores, the models performed worse on this set compared to the balanced non-augmented set. The ANN, KNN, Logistic Regression, and Gradient Boosting models each had slightly higher accuracy when fit on this augmented training set. These results also support the necessity of examining multiple performance metrics, as each of these models had relatively low precision and sensitivity scores. Figure 10 summarizes each model's performance results and their respective ROC curves.

Figure 10: Performance Summary & ROC Curves for Classifier Models - Balanced Additive Noise Data

Summary of Results for Balanced Augmented Data

Model Performance Results - Balanced Augmented Training Data # 1 (Additive Noise)

		DISEASE CLASS = EFFUSION	Precision	Sensitivity/Recall	F1-Score	Accuracy	AUC
Supervised Learning	Classification	ANN	0.529709	0.560160	0.544509	0.704593	0.66
		KNN	0.338812	0.663366	0.448536	0.722292	0.63
		kernel-SVM	0.355247	0.667458	0.463696	0.726085	0.63
		Logistic Regression	0.474083	0.554734	0.511247	0.697851	0.64
		Naïve Bayes	0.699115	0.502727	0.584876	0.669195	0.68
		Gradient Boosting	0.37547	0.65708	0.477876	0.726507	0.64
Unsupervised Learning	Clustering (Classification)	K-means	0.70291	0.346633	0.464301	0.459334	0.52
	Dimensionality Reduction	PCA	*No dimensionality reduction was performed on the data used to fit and test these models.*				



Balanced Augmented Data Set - Blurring

The balanced augmented data with blurring consisted of a total of 15,820 images, 80% of which were used for training and the remaining 20% for testing. The seven classification models were trained on 12,656 images and then tested on 3,164 images. Of the seven models tested, the Naïve Bayes classifier and the Artificial Neural Network classifier had the best performance with AUC scores of 0.68 and 0.67 and test accuracies of 67.0% and 71.4%, respectively. Both of

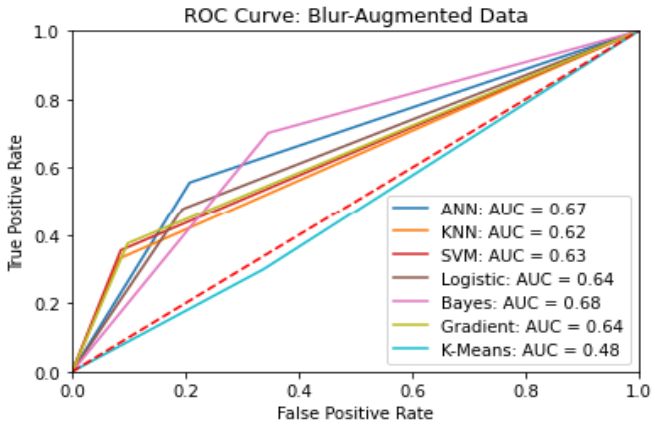
these models notably had comparable accuracy and AUC scores than the Naïve Bayes and ANN models fit on the balanced non-augmented data set, though their F-1 scores for this augmented data set were relatively lower. Figure 11 summarizes each model's performance results and their respective ROC curves.

Figure 11: Performance Summary & ROC Curves for Classifier Models - Balanced Blurred Data

Summary of Results for Balanced Augmented Data

Model Performance Results - Balanced Augmented Training Data # 2 (Blurring)

		Disease Class = Effusion	Precision	Sensitivity/Recall	F1-Score	Accuracy	AUC
Supervised Learning	Classification	ANN	0.556258	0.572917	0.564464	0.713864	0.67
		KNN	0.331226	0.663291	0.441821	0.721028	0.62
		kernel-SVM	0.355247	0.673861	0.465232	0.727771	0.63
		Logistic Regression	0.477876	0.551020	0.511848	0.696165	0.64
		Naïve Bayes	0.701643	0.503630	0.586371	0.670038	0.68
		Gradient Boosting	0.375474	0.65708	0.477876	0.726507	0.64
Unsupervised Learning	Clustering (Classification)	K-means	0.299621	0.306995	0.303263	0.541087	0.48
	Dimensionality Reduction	PCA	*No dimensionality reduction was performed on the data used to fit and test these models.*				



Balanced Augmented Data Set - Cropping

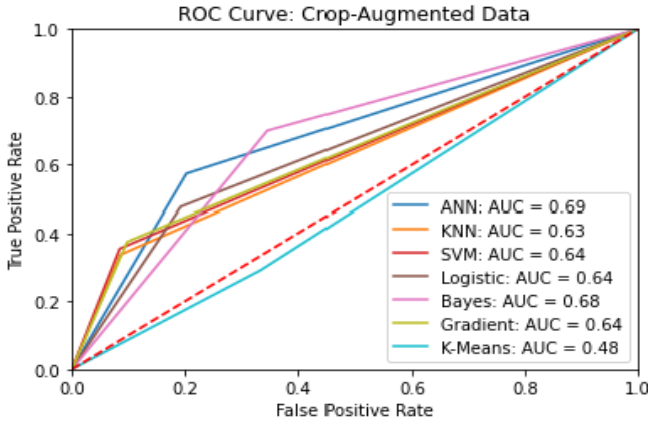
The balanced augmented data with cropping consisted of a total of 15,820 images, 80% of which were used for training and the remaining 20% for testing. The seven classification models were trained on 12,656 images and then tested on 3,164 images. Of the seven models tested, the Artificial Neural Network (ANN) classifier and the Naïve Bayes classifier had the best performance with AUC scores of 0.69 and 0.68, respectively; and model test accuracies of 72.3% and 66.9%, respectively. The findings here are similar to the previous augmented data set, as each of the ANN and Naïve Bayes models had comparable accuracy and AUC scores to the balanced non-augmented data set, but much lower F1-scores. Figure 12 summarizes each model's performance results and their respective ROC curves.

Figure 12: Performance Summary & ROC Curves of Classifier Models - Balanced Cropped Data

Summary of Results for Balanced Augmented Data

Model Performance Results - Balanced Augmented Training Data # 3 (Cropping)

		Disease Class = Effusion	Precision	Sensitivity/Recall	F1-Score	Accuracy	AUC
Supervised Learning	Classification	ANN	0.573957	0.585806	0.579821	0.722714	0.69
		KNN	0.337547	0.662531	0.447236	0.721871	0.63
		kernel-SVM	0.355247	0.677108	0.466003	0.728614	0.64
		Logistic Regression	0.476612	0.552786	0.511881	0.697008	0.64
		Naïve Bayes	0.700379	0.502722	0.585314	0.669195	0.68
		Gradient Boosting	0.375474	0.65708	0.477876	0.726507	0.64
Unsupervised Learning	Clustering (Classification)	K-means	0.292035	0.303947	0.297872	0.541087	0.48
	Dimensionality Reduction	PCA	*No dimensionality reduction was performed on the data used to fit and test these models.*				



Balanced Augmented Data Set - Horizontal and Vertical Flipping

The balanced augmented data with horizontal and vertical reflection geometric transformations consisted of a total of 15,820 images, 80% of which were used for training and the remaining 20% for testing. The seven classification models were trained on 12,656 images and then tested on 3,164 images. Of the seven models tested, the Artificial Neural Network (ANN) classifier and the Naïve Bayes classifier had the best performance with AUC scores of 0.70 and 0.68, respectively; and model test accuracies of 74.5% and 66.3%, respectively. This is the best accuracy and AUC score achieved by the ANN model yet, though notably it still suffers from the

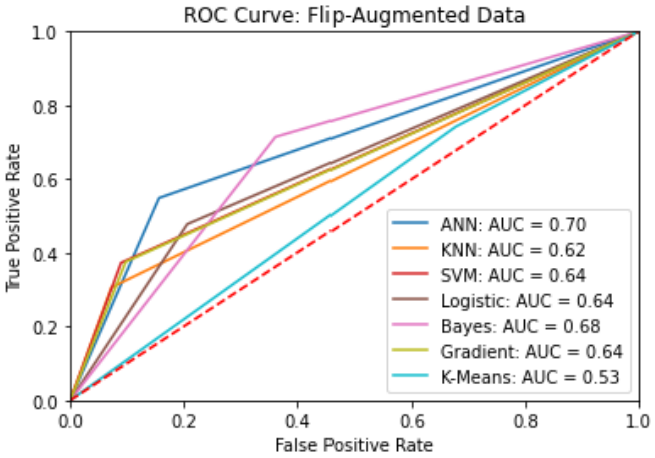
same issue of previously examined augmented data sets as the model's F-1 score is still rather low. Figure 13 summarizes each model's performance results and their respective ROC curves.

Figure 13: Performance Summary & ROC Curves of Classifier Models - Balanced Flipped Data

Summary of Results for Balanced Augmented Data

Model Performance Results - Balanced Augmented Training Data # 4 (Flipping)

DISEASE CLASS = EFFUSION			Precision	Sensitivity/Recall	F1-Score	Accuracy	AUC
Supervised Learning	Classification	ANN	0.548673	0.636364	0.589274	0.745048	0.70
		KNN	0.305942	0.677871	0.421603	0.720185	0.62
		kernel-SVM	0.372946	0.675057	0.480456	0.731142	0.64
		Logistic Regression	0.477876	0.536170	0.505348	0.688158	0.64
		Naïve Bayes	0.714286	0.496485	0.585796	0.663295	0.68
		Gradient Boosting	0.37547	0.65708	0.477876	0.726507	0.64
Unsupervised Learning	Clustering (Classification)	K-means	0.7421	0.353828	0.479184	0.462284	0.53
	Dimensionality Reduction	PCA	*No dimensionality reduction was performed on the data used to fit and test these models.*				



Discussion

Model Performance

Expectedly, the classification models trained on the imbalanced, non-augmented dataset performed the worst across all combinations of dataset balancing and data augmentation methods when using AUC scores as the primary aggregate performance metric. These results align with the tendency for classification models to struggle when presented with incongruous class distributions (*Sun et al., 2009*). When the training set consists of mostly one class, the algorithm fails to learn the differences between classes and develops a bias towards the majority class. The models' high accuracy values, despite low AUCs, are another manifestation of this. The accuracy scores are inflated from correctly predicting the majority class, regardless of its inability to classify the minority class or distinguish the two. The Naive Bayes model, however, performed similarly to other balanced and augmented data sets, perhaps due to its "naive" class independence assumption that features are independent, and its ability to better fit smaller datasets compared to more sophisticated models.

With an AUC of 0.75 and an accuracy of 76%, the kernel-SVM model trained on balanced, non-augmented data was the best overall model tested. This partially answers the first research question, given the understanding that the kernel-SVM model trained on balanced, non-augmented data performed the best across all the models and data sets used. While the accuracy value is one of the most common and straightforward metrics measuring the percentage of correct prediction by the model, the AUC value plots the relation of the true positive rate and false positive rate and is an effective measurement for comparing performance across multiple models. We can also trust our accuracy metric more heavily when evaluating models trained on the balanced dataset compared to the imbalance dataset, as explained previously. We can observe how kernel-SVM is an effective image classification algorithm, given its ability to overcome linearly inseparable data and generate a decision boundary to classify the images. The result of tuning this model was the use of a polynomial kernel function with a third-degree polynomial, which partially explains this model's relative success with classifying non-linearly separable data given its choice of a decision boundary. This explanation helps answer the second research question as to what the kernel-SVM model tracks and measures can be explained when considering how its decision boundary is formed and its related ROC curve.

None of the data augmentation methods and related models yielded better results than our baseline of the balanced, non-augmented dataset, suggesting that our original non-augmented models were overfitted or that our dataset is still too small for our models to effectively learn

from. Additionally, this partially answers the third research question as it can be seen that each augmentation technique had a negative effect on the best performing model, the kernel-SVM model. With an AUC score of 0.70 and accuracy of 74.5%, however, the ANN model trained on balanced, flip-augmented data performed the closest to our balanced, non-augmented kernel-SVM model. Given that a neural network must estimate a larger number of parameters compared to the other models tested in this study, it would be assumed that the use of larger data sets would benefit the ANN models considerably. With this understanding, given that a reasonable data augmentation technique can be used to generate more training data, it would be assumed that the ANN model could see an improvement in performance, and such is the case here in regard to the balanced, flip-augmented dataset. Additionally, the Naive Bayes models scored consistently higher than most other models, across all augmentation methods. It would appear that the underlying assumptions of this model make it less sensitive to changes in features in the data set, causing its performance to vary less when effective data augmentation techniques are used, or not. Furthermore, when comparing the performance of all models across each augmented data set, the crop-augmented data yielded consistently better results than other augmentation methods. This suggests that cropped images were more relevant to models' mapping features compared to image blur, additive noise, and flipping, and could perhaps be a viable augmentation technique for increasing data set size when combined with other effective augmentation methods. Additionally, this helps answer the third research question given the understanding that cropping was the most effective augmentation technique tested.

Comparison to Deep Learning Models

Most models generated within this methodology did not perform reasonably compared to previous deep learning models applied to the same dataset. The kernel-SVM model trained using the balanced, non-augmented dataset did perform relatively comparably to previous deep learning models, however, with an AUC of 0.75. Most notably, this kernel-SVM model was able to be trained and tuned at a fraction of the computational cost compared to similar deep learning models, and with the added benefit of interpretability that deep learning models do not possess.

Other models also performed relatively well, most notably, the KNN and gradient boosting models generated from the same dataset which performed similarly with AUC's of 0.72. A full comparison of the models fit in this study and relevant deep learning models applied on the same dataset can be seen in Figure 14.

Figure 14: Previous DL Results on the CXR data set for Disease Classification

		Effusion AUC Scores for Interpretable ML Models Used in This Work, Compared to the Results of Previous Work that Implemented DL Models (Chen, et al., 2022)												
		AUC for ML Models in this work							AUC for Previously Reported Deep Learning (DL) Models					
		ANN	KNN	k-SVM	Logistic Regression	Naïve Bayes	Gradient Boosting	K-Means	Chen[15]	Huang[33]	Wang[28]	Yao[32]	Baltruchat[9]	Li[16]	Wang[34]
AUC scores	Imbalanced, Non-Augmented	0.52	0.5	0.5	0.52	0.68	0.5	0.51	0.879	0.839	0.818	0.806	0.818	0.789	0.759
	Balanced, Non-Augmented	0.69	0.72	0.75	0.62	0.69	0.72	0.55							
	Balanced, Noise-Augmented	0.66	0.63	0.63	0.64	0.68	0.64	0.52							
	Balanced, Blur-Augmented	0.67	0.62	0.63	0.64	0.68	0.64	0.48							
	Balanced, Crop-Augmented	0.69	0.63	0.64	0.64	0.68	0.64	0.48							
	Balanced, Flip-Augmented	0.70	0.62	0.64	0.64	0.68	0.64	0.53							

This shows the tradeoff between model performance and interpretability when fitting models for medical image analysis. While deep learning models may perform better than the models tested in this study, they are not interpretable and therefore are not viable for clinical implementation. Given this understanding, one could argue that opting for simpler, more interpretable models which perform comparably should be implemented over deep learning models.

Limitations

There are multiple limitations to consider before drawing conclusions from these research questions. Primarily, despite having more than 100,000 images, it was not apparent how imbalanced the data was until performing deeper data exploration through initial data processing. Although downsampling the dataset was a feasible solution to move forward with analysis, this reintroduced issues pertaining to small datasets.

Time and computational resource constraints also limited how extensively we were able to explore these concepts; notably, parameter tuning via random combinations rather than parsing, evaluating, and comparing every combination of model parameters. Without time and computational barriers, it would have been more viable to use other augmentation methods, such as Generative Adversarial Networks (GANs). Although GANs can create impressive, entirely new samples that resemble the feature distribution of the original images, they are quite computationally costly to generate. Augmenting through GANs would create new challenges like labeling synthetic data, but unfortunately, we were not able to thoroughly explore that augmentation method.

Conclusions and Recommendations for Future Research

A medical image analysis framework was designed and implemented to address the concern of black-box models in healthcare. Variations of interpretable machine learning models were trained and tested on the NIH Chest X-ray image data set for classifying the pathology of "Effusion" versus "No Finding". In addition, data augmentation techniques were utilized to address the issue of limited training data. Seven machine learning models, ANN, KNN, kernel-SVM, Logistic Regression, Naïve Bayes, Gradient Boosting, and K-Means, were tuned, trained, and tested for each data set variation. These variations included balanced non-augmented data, imbalanced non-augmented data, and balanced augmented data using additive noise, blurring, cropping, or flipping image transformations for augmentation. The results of this work indicate that ANN and Naïve Bayes performed best when using augmented data sets with AUC scores between 0.67 and 0.70, and kernel-SVM performed best overall when using the non-augmented balanced data with an AUC score of 0.75. Notably, this kernel-SVM model demonstrated a slightly worse, but comparable performance to the deep learning models reported in previous works (*Chen et al., 2022*).

The impact of imbalanced data on classifier performance was remarkable. With the balanced non-augmented data set, the kernel-SVM model performed best with an AUC score of 0.75. The same kernel-SVM classifier applied to imbalanced data was not able to distinguish between a pathology finding of "Effusion" and "No Finding". Augmentation of the training data yielded some improvement for the kernel-SVM classifier, with AUC scores of 0.63-0.64. This profound impact of imbalanced data on model performance provides supporting evidence that there is a need for balanced data to achieve the task of using interpretable machine learning models for classifying lung disease from chest X-ray images.

Data augmentation did not improve performance, however, this could potentially be re-evaluated by improving upon the methods tested in this work. For example, future work could try combining various augmentation methods and testing varying levels of augmentation: no augmentation, simple augmentation, and complex augmentation. Most notably, more complex augmentation techniques such as Generative Adversarial Networks could be utilized to create new data of a much higher quality than the basic augmentation techniques utilized in this study (*Suganyadevi et al., 2022*). In addition, it could be beneficial to have clinicians and radiologists review the augmented data for their expert opinions on whether or not the augmented images are representative samples of clinical findings in real-world medical imaging. Lastly, model tuning parameters may need further exploration for handling a wider range of pathologies

and/or image augmentation types. Future work could try re-tuning models for augmented training data to achieve better model performance.

The results of this study suggest a possibly comparable and interpretable machine learning model for medical image classification which addresses issues of opacity and plasticity inherent to deep learning models used for the same task, and highlights the effect of data set balancing and augmentation for model performance.

References

1. Alarifi, M., Patrick, T., Jabour, A., Wu, M., & Luo, J. (2021). Understanding patient needs and gaps in radiology reports through online discussion forum analysis. Insights Imaging, 12(1), 50. https://doi.org/10.1186/s13244-020-00930-2

2. Apostolopoulos, I. D., & Mpesiana, T. A. (2020). Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine, 43(2), 635-640. https://doi.org/10.1007/s13246-020-00865-4

3. Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., & Saalbach, A. (2019). Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. Scientific Reports, 9(1), 6381. https://doi.org/10.1038/s41598-019-42294-8

4. Bengio, Y. (2012). Practical Recommendations for Gradient-Based Training of Deep Architectures. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), Neural Networks: Tricks of the Trade: Second Edition (pp. 437-478). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35289-8_26

5. Chen, H., Gomez, C., Huang, C.-M., & Unberath, M. (2022). Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. npj Digital Medicine, 5(1), 156. https://doi.org/10.1038/s41746-022-00699-2

6. Chen, K., Wang, X., & Zhang, S. (2022). Thorax Disease Classification Based on Pyramidal Convolution Shuffle Attention Neural Network. IEEE Access, 10, 85571-85581. https://doi.org/10.1109/ACCESS.2022.3198958

7. Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology, 65(5), 545-563. https://doi.org/10.1111/1754-9485.13261

8. Chollet, F. (2015). Keras. In https://github.com/fchollet/keras

9. Civit-Masot, J., Luna-Perejón, F., Domínguez Morales, M., & Civit, A. (2020). Deep Learning System for COVID-19 Diagnosis Aid Using X-ray Pulmonary Images. Applied Sciences, 10(13), 4640. https://www.mdpi.com/2076-3417/10/13/4640

10. Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. J Med Ethics, 46(3), 205-211. https://doi.org/10.1136/medethics-2019-105586

11. Hassan, M., Ali, S., Alquhayz, H., & Safdar, K. (2020). Developing intelligent medical image modality classification system using deep transfer learning and LDA. Scientific Reports, 10(1), 12868. https://doi.org/10.1038/s41598-020-69813-2

12. Holste, G., Wang, S., Jiang, Z., Shen, T. C., Shih, G., Summers, R. M., Peng, Y., & Wang, Z. (2022). Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study. Data Augment Label Imperfections (2022), 13567, 22-32. https://doi.org/10.1007/978-3-031-17027-0_3

13. Huang, Z., Lin, J., Xu, L., Wang, H., Bai, T., Pang, Y., & Meen, T. H. (2020). Fusion high-resolution network for diagnosing ChestX-ray images. *Electronics*, *9*(1), 190.

14. Kondylakis, H., Ciarrocchi, E., Cerda-Alberich, L., Chouvarda, I., Fromont, L. A., Garcia-Aznar, J. M., Kalokyri, V., Kosvyra, A., Walker, D., Yang, G., & Neri, E. (2022). Position of the AI for Health Imaging (AI4HI) network on metadata models for imaging biobanks. Eur Radiol Exp, 6(1), 29. https://doi.org/10.1186/s41747-022-00281-1

15. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

16. Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L. J., & Fei-Fei, L. (2018). Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8290-8299).

17. Marias, K. (2021). The Constantly Evolving Role of Medical Image Processing in Oncology: From Traditional Medical Image Processing to Imaging Biomarkers and Radiomics. Journal of Imaging, 7(8), 124. https://www.mdpi.com/2313-433X/7/8/124

18. Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., & Jamalipour Soufi, G. (2020). Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. Medical Image Analysis, 65, 101794. https://doi.org/10.1016/j.media.2020.101794

19. Ng, H. P., Ong, S. H., Foong, K. W. C., Goh, P. S., & Nowinski, W. L. (2006, 26-28 March 2006). Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm. 2006 IEEE Southwest Symposium on Image Analysis and Interpretation, 61-65. DOI: 10.1109/SSIAI.2006.1633722

20. Panwar, H., Gupta, P. K., Siddiqui, M. K., Morales-Menendez, R., & Singh, V. (2020). Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. Chaos, Solitons & Fractals, 138, 109944. https://doi.org/10.1016/j.chaos.2020.109944

21. Price, W. N. (2018). Big data and black-box medical algorithms. Science Translational Medicine, 10(471), eaao5333. https://doi.org/doi:10.1126/scitranslmed.aao5333

22. Puttagunta, M., & Ravi, S. (2021). Medical image analysis based on deep learning approach. Multimedia Tools and Applications, 80(16), 24365-24398. https://doi.org/10.1007/s11042-021-10707-4

23. Shad, R., Cunningham, J. P., Ashley, E. A., Langlotz, C. P., & Hiesinger, W. (2021). Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging. Nature Machine Intelligence, 3(11), 929-935. https://doi.org/10.1038/s42256-021-00399-8

24. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(1), 60. https://doi.org/10.1186/s40537-019-0197-0

25. Shrivastava, K., Gupta, N., & Sharma, N. (2014). Medical Image Segmentation using Modified K Means Clustering. International Journal of Computer Applications, 103, 12-16. DOI: 10.5120/18157-9341

26. Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2022). A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval, 11(1), 19-38. https://doi.org/10.1007/s13735-021-00218-1

27. Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(4), 867-719. https://doi.org/10.1142/S0218001409007326

28. Team, T. T. D., Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., & Belikov, A. (2016). Theano: A Python framework for fast computation of mathematical expressions. arXiv preprint arXiv:1605.02688.

29. van Leeuwen, K. G., de Rooij, M., Schalekamp, S., van Ginneken, B., & Rutten, M. J. C. M. (2022). How does artificial intelligence in radiology improve efficiency and health outcomes? Pediatric Radiology, 52(11), 2087-2093. https://doi.org/10.1007/s00247-021-05114-8

30. Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digital Medicine, 5(1), 48. https://doi.org/10.1038/s41746-022-00592-y

31. Wang, H., Jia, H., Lu, L., & Xia, Y. (2019). Thorax-net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography. *IEEE journal of biomedical and health informatics*, *24*(2), 475-485.

32. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3462-3471.

33. Yao, L., Prosky, J., Poblenz, E., Covington, B., & Lyman, K. (2018). Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*.

34. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683

Acknowledgements

We acknowledge and thank Ronald M. Summers, M.D., Ph.D. (https://www.cc.nih.gov/drd/summers.html), Senior Investigator of the Clinical Image Processing Service in the Imaging Biomarkers and Computer-Aided Diagnosis Laboratory of the NIH Clinical Center Radiology and Imaging Sciences Department. Dr. Summers' work, supported by the NIH Clinical Center (clinicalcenter.nih.gov) and National Library of Medicine (www.nlm.nih.gov), provides one of the largest publicly available chest X-ray datasets to the scientific community. In addition, we thank Georgia Institute of Technology for access to software and content covered in Computational Data Analysis (Machine Learning), that made this work possible.