Trustworthiness of Tuberculosis Chest X-ray Classification Models

Linsey Szabo, Desmond Kuan, and Haineng Huang (Clark)

***Abstract*— With the rise of artificial intelligence (AI), it is easy to see that if AI is successfully able to aid physicians, they can focus on more complex work and overall increase their efficiency. However, in the AI diagnostic tool space, physicians lack trust in these tools [1], and as a result they are unwilling to use their recommendations. We compare AlexNet and InceptionV3 as tuberculosis classification models for chest X-rays. To explore the robustness of these models for the purpose of building towards a benchmark for trustworthiness in AI diagnostic tools, we record accuracy and F1 metrics on noisy images and out-of-distribution images, specifically a COVID-19 chest X-ray dataset.**

***Clinical Relevance*— This is a brief statement on why a this might be of interest to practicing clinicians.** Example**: This establishes the anesthetic efficacy of 10% intraosseous injections with epinephrine to positively influence cardiovascular function. Maximum of 50 words.**

I. Introduction

There is currently a need for trustworthy AI in the biomedical engineering space. These models are not only dealing with sensitive data, but they are also deployed in sometimes life-threatening situations. Therefore, it is paramount that physicians can trust the AI that aids them in their work.

More specifically, X-rays provide interesting challenges for AI models. Due to different X-ray technicians, there are variations in the resulting images [2]. These differences can be hard for an AI model to generalize about. Furthermore, for humans, X-ray images are hard to read. The goal of an AI diagnostic model would ultimately be to point out diagnoses doctors hadn't considered before and, in this way, facilitate early diagnosis, which is key to quicker healing timelines.

Delving deeper into the differences X-ray machines may produce, there are also additional sources of noise for X-ray images in practice [2]. A fidgety patient produces a slightly blurry image; a patient may forget to remove their jewelry or stick-on sensor which produces a clear artifact in the image; a technician takes an image that is slightly off-axis. Specific to using these X-ray images in an AI model, patient-specific annotations must be covered with white boxes. We aim to simulate these sources like these using Python packages like OpenCV. Then we observe the corresponding accuracy and F1 score as noise is increased over 6 levels. In this way, we can discuss robustness of these models to better understand how we can trust AI models as diagnostic tools.

In addition to image degradation, we also examine model performance on a COVID-19 chest X-ray dataset. We chose this as an out-of-distribution dataset. A doctor when evaluating an X-ray image considers multiple diagnoses. However, an AI binary classification model has clear limitations on what it can classify. By testing on COVID-19 data, we hoped to explore what might happen when a model is exposed to out-of-distribution data.

Finally, we used two models to perform our analyses. For a baseline CNN model, we chose AlexNet, which features the classical convolutional architecture with max pooling. For a state-of-the-art image classification model, we chose InceptionV3, which has many more convolutional layers. We want to examine how models of different complexity handle degraded X-ray images and out-of-distribution data.

II. Methods

A. Dataset Acquisition

A tuberculosis (TB) chest X-ray database found on Kaggle [3] was utilized to train the model and generate our novel image degradation dataset for testing. It consists of 3,500 normal images and 700 TB images. Additional TB images were accessed from the National Institute of Allergy and Infectious Diseases (NIAID) TB portal. Due to data sharing restrictions, our team signed an NIAID data sharing agreement, co-signed by our course instructor, which granted us access to the Digital Imaging and Communications in Medicine (DICOM) images directly. We downloaded the TB chest X-ray DICOM files from the portal, converted them to the PNG format, and added them to our training dataset, resulting in a more balanced dataset with 3,500 Normal and 2,455 TB samples.

To assess the robustness of our models on out-of-distribution data, we used a COVID Database from Kaggle, which encompasses 3,500 normal and 3,500 COVID cases [4].

B. Data Preprocessing

Per similar experiments who had used the same TB dataset [5], we used the ImageDataGenerator function from Keras' image processing package [6]. In this way, we could perform online data augmentation during training only. Some data transformations include rescaling all input images to the same resolution of (512, 512), rescaling all pixels by 1/255, and slight horizontal and vertical shifts of the images.

C. Model Selection

We opted for AlexNet [7] and Inception V3 [8]. AlexNet was chosen due to its proven performance in image classification tasks, its simplicity, and faster training compared to deeper networks [7]. It serves as a baseline model for comparison in our project. Inception V3 was selected for its state-of-the-art performance, efficient use of computational resources, and robustness to variations in image quality and

scale [8]. This model provides a more complex alternative for comparison with our baseline, AlexNet.

D. AlexNet

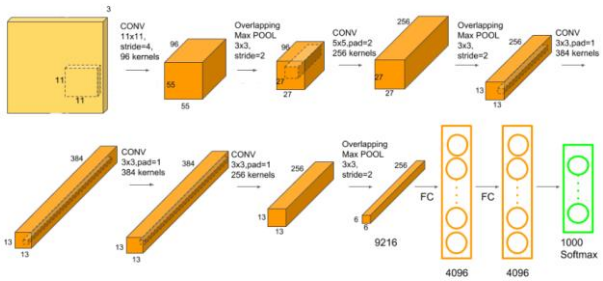

Figure 1: AlexNet architecture [7]

For AlexNet, the original implementation [9] was followed, except for the normalization method. In the first two layers, batch normalization was applied to our baseline model instead of local response normalization. Batch normalization helps to normalize the activations of the previous layer, improving the training process by reducing internal covariate shift and allowing the use of higher learning rates. This technique has been proven to enhance model generalization and accelerate training [10]. No batch normalization is applied for the remaining layers.

E. Inception V3

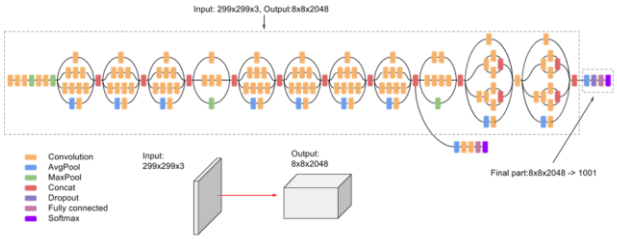

Figure 2: Inception V3 architecture [11]

Using the example of Kaggle [5], the Inception V3 model can be imported through the Keras API. We wanted to use the pretrained model weights but train the output layers specifically for the tuberculosis classification problem. This was able to be accomplished by marking the weights of the pretrained layers as untrainable. By adding two fully connected layers with dropout followed by a sigmoid activation function, we finished the Inception V3 model for our study.

F. Image Degradation

For data degradation, we have 6 classes and a total of 16 approaches:

- We noticed that it's common for data annotators to cover the patient's name on the X-ray images with white or black boxes for privacy protection, but we assume that may lead to a decrease in model prediction accuracy, so we use the Python Pillow package to add different sizes and numbers of white boxes on the testing pictures across different levels. Note that to better simulate the real-world situation, the boxes are not in any center area of the pictures in a [25%,75%] scale.

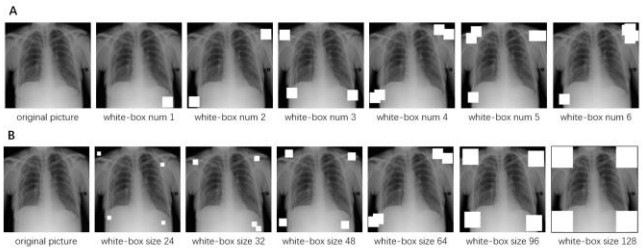

Figure 3: Original picture and different degradation levels with adding boxes and changing box size

- The second degradation attempted to simulate the image quality under different X-ray intensities. An ideal way to simulate this problem is to use many kinds of image quality changing methods (brightness, contrast, saturation, and hue) at the same time and find the best weight for different levels. However, due to the lack of a dataset to provide a benchmark for these weights, we are only going to change only one parameter each time and compare them separately.

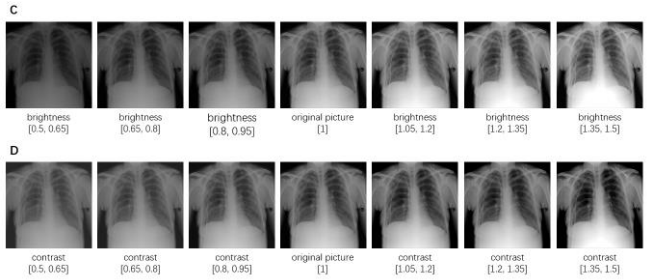

Figure 4: Original picture and different degradation levels for brightness and contrast

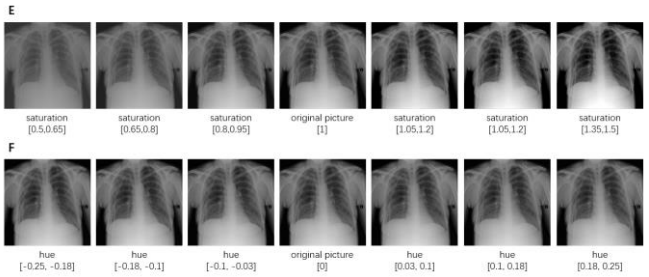

Figure 5: Original picture and different degradation levels for saturation and hue

- Backscatter is another effect that arises from scattered radiation reflecting off the back of the detector. It produces an image of the electronic components of the detector overlaid on the patient image, and it is common with portable equipment and high-dose radiation [12]. We use the Python OpenCV package to overlay the original images

and backscatter mask with a certain weight as degradation.

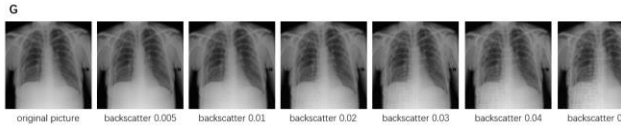

Figure 6: Original picture and different degradation levels for backscatter overlay weight

- Another kind of artifact in X-ray imaging is blurring or noise, which include Poisson noise, Gaussian noise, and the motion blurring. The motion blurring is more common in the real world, which is usually caused by the subject's movement or double exposure of the system. Here, we choose Gaussian noise as a baseline to compare with and focus on motion blurring.

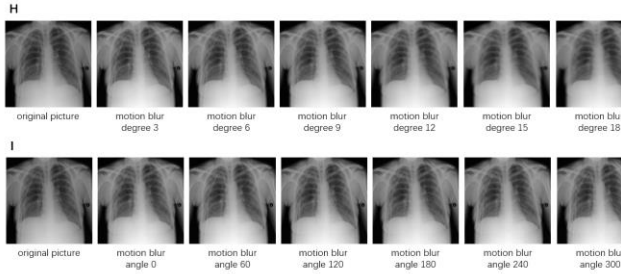

Figure 7: Original picture and different degradation levels for motion blurring

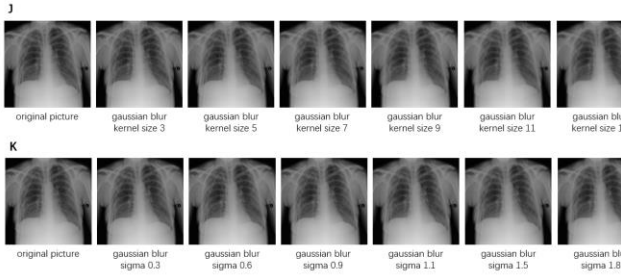

Figure 8: Original picture and different degradation levels for Gaussian blurring

- According to a survey on radiographers [2], the most common reason why they would regard an X-ray image as unfit is suboptimal positioning and centering, so we used tensorflow.transforms.random_perspective and tensorflow.transforms.rotation to simulate different perspectives and rotations.

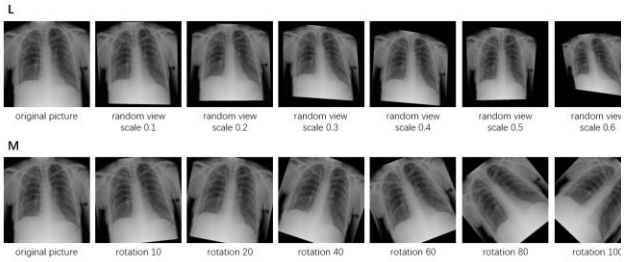

Figure 9: Original picture and different degradation levels for perspective and rotation

- The last degradation approach is adding objects. Checking our dataset, we found 3 kinds of object artifacts (chest detector, necklace, and collar protector), so we picked 5 pictures of each and did Photoshop to get the artifact area itself in new PNG files and overlaid them with the original images as degradation. Note that this degradation does not have different levels.

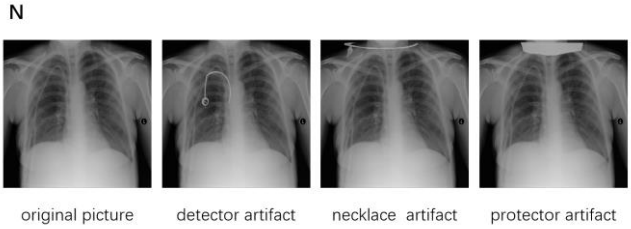

Figure 10: Original picture and different degradation for artifacts

G. *Testing*

After we finished training both models on the original dataset, we tested them on the degradation data to get their performance (accuracy and F1 score) under different levels of degradation. Note that the models are only trained with the original dataset; no degradation data is used for training.

III. RESULTS

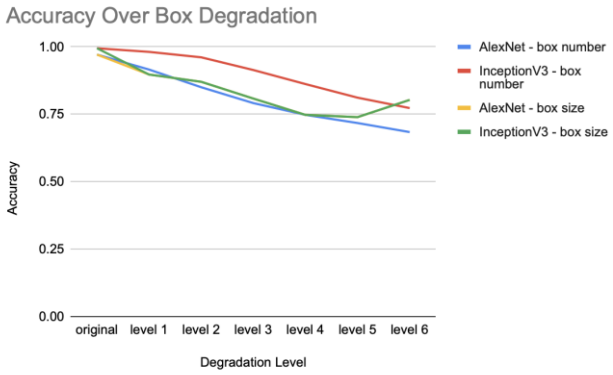

Figure 11: Accuracy over box degradation

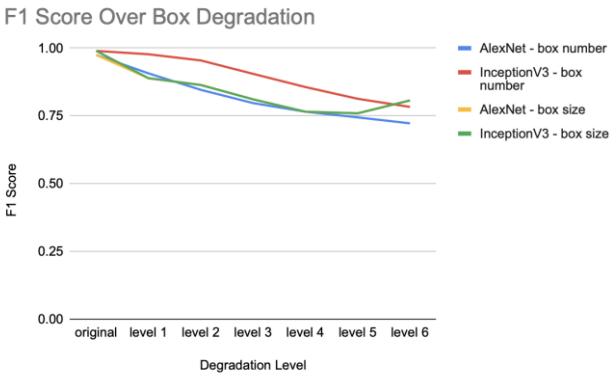

Figure 12: F1 score over box degradation

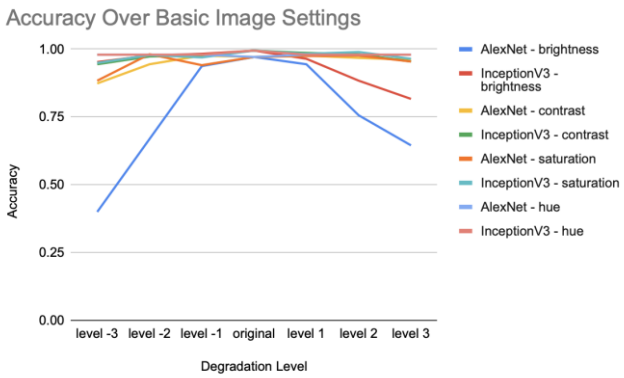

Figure 13: Accuracy over basic image settings

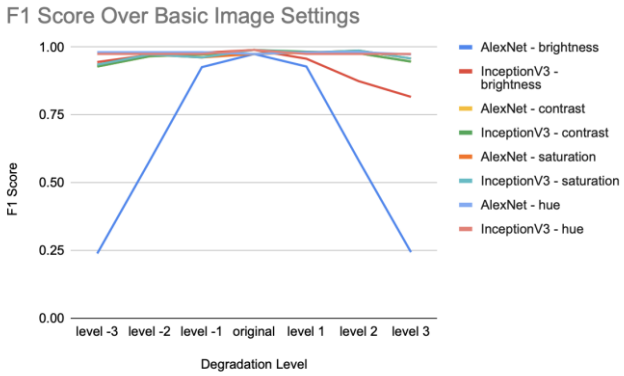

Figure 14: F1 score over basic image settings

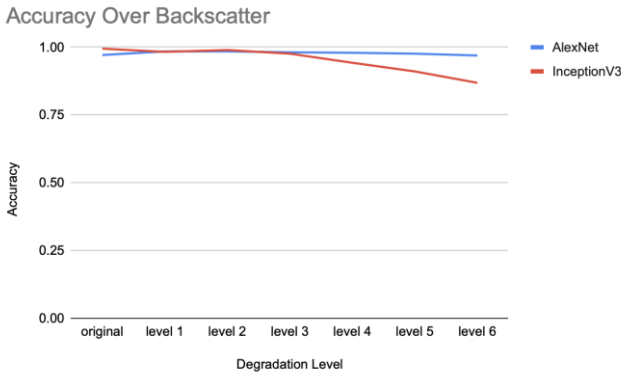

Figure 15: Accuracy over backscatter

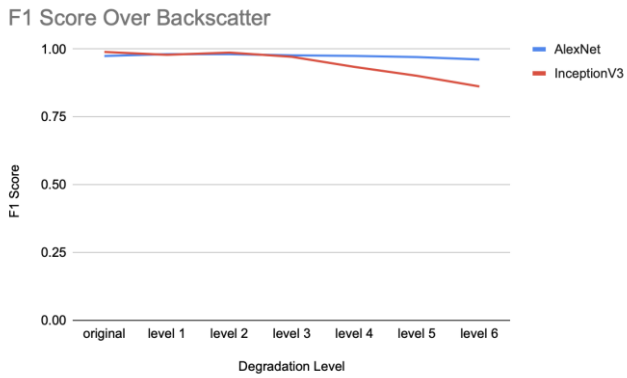

Figure 16: F1 score over backscatter

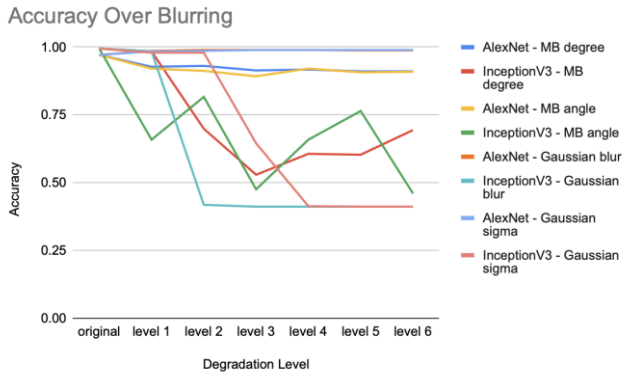

Figure 17: Accuracy over blurring

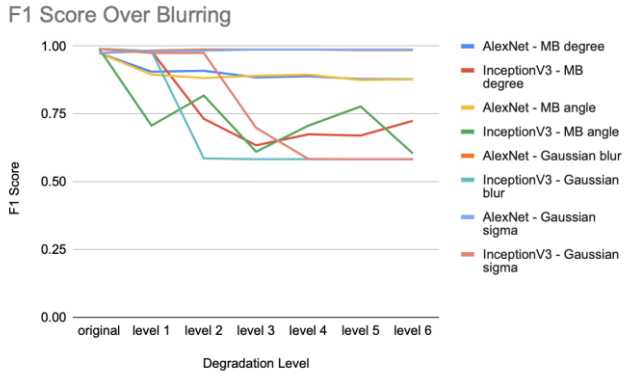

Figure 18: F1 score over blurring

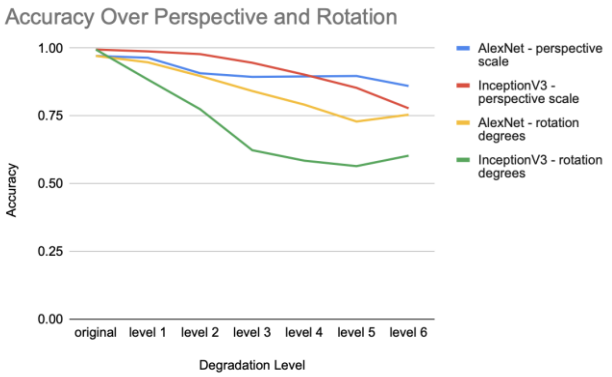

Figure 19: Accuracy over perspective and rotation

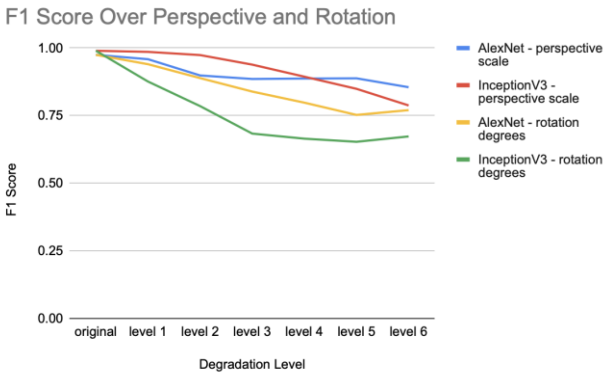

Figure 20: F1 score over perspective and rotation

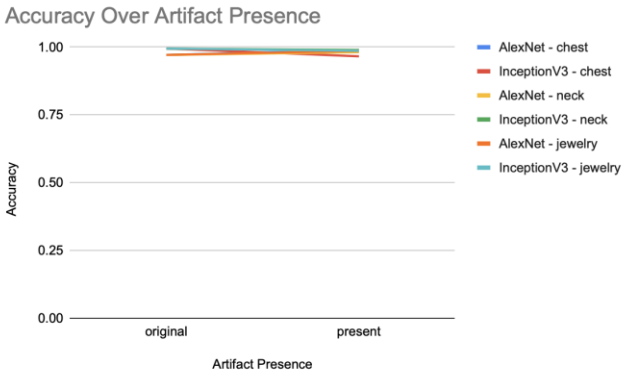

Figure 21: Accuracy over artifact presence

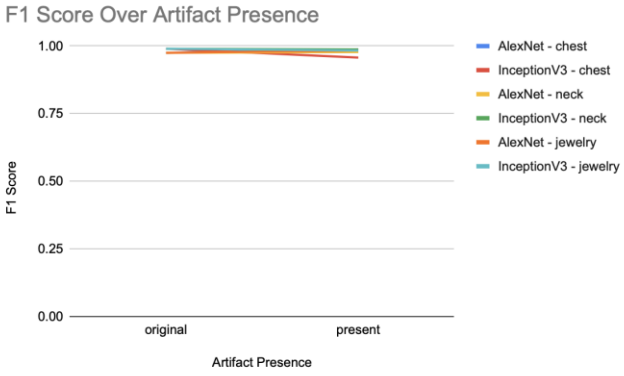

Figure 22: F1 score over artifact presence

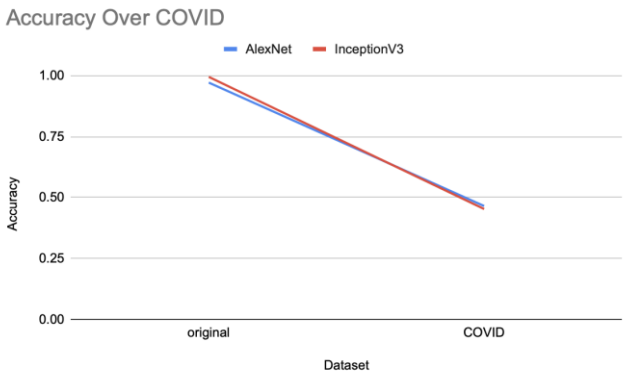

Figure 23: Accuracy for COVID dataset

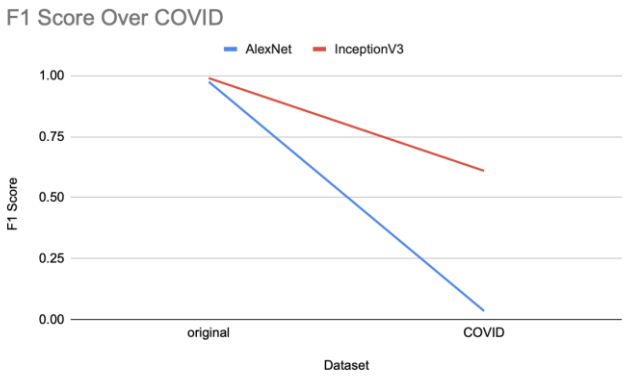

Figure 24: F1 score for COVID dataset

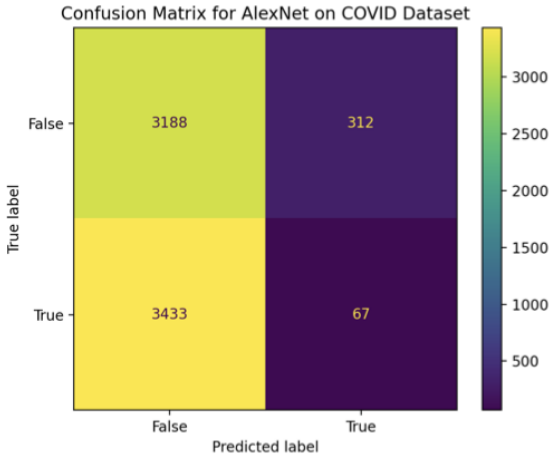

Figure 25: Confusion matrix for AlexNet on COVID dataset

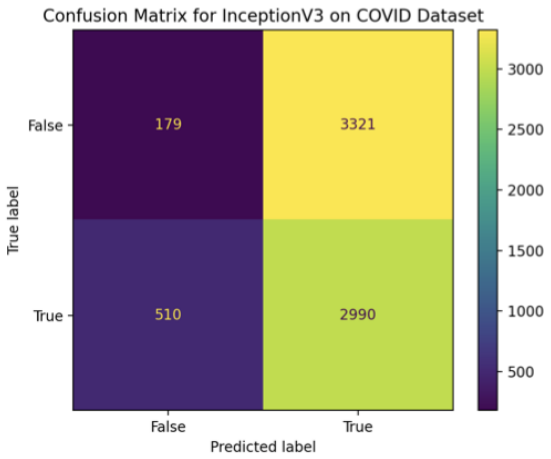

Figure 26: Confusion matrix for Inception V3 on COVID dataset

IV. Discussion

A. *Image Degradation*

1. Box degradation

While Inception V3 does outperform AlexNet, across both networks there is a relatively low drop in both accuracy and F1 score as more boxes are added and box size increases. From level 5 to level 6, there is a notable uptick in accuracy and F1 score for Inception V3 and box size. We hypothesize that this is due to random chance.

2. Basic image settings

AlexNet suffers a dramatic decrease in performance as brightness is significantly increased or decreased. InceptionV3's accuracy and F1 score also branch down for increased levels of brightness. We can infer that regardless of model complexity they can suffer to recognize features of images with which to identify them (the strength of CNN's) under different brightness levels.

3. Backscatter

AlexNet outperforms Inception V3. We hypothesize that due to the complexity of Inception V3 its strength should lie in extract features from these X-ray images. However, because we overlay a mask of the electronic components of the detector, it may in fact be detecting this mask instead of the X-ray image itself.

4. Blurring

We observe the lowest accuracies and F1 scores for any image degradation effect for Gaussian blur. However, we do focus on the motion blur degradation, which also has quite low accuracy and F1 scores for both networks, pointing to the important of this degradation effect and image classification.

5. Perspective and rotation

Inception V3 does the worst with respect to rotation. This points to another weakness in this state-of-the-art architecture. Rotation decreases accuracy and F1 score more than perspective for both networks.

6. Artifact presence

None of the added artifacts affected the performance of either network.

B. *Out-of-distribution data*

We can see that Inception V3 classified most of the COVID-19 dataset as tuberculosis; AlexNet classified most of the dataset as normal. In considering how neural networks arrive at their decision boundaries, the decision boundary of the Inception V3 network classified COVID very similarly to tuberculosis, while AlexNet classified COVID very similarly as normal.

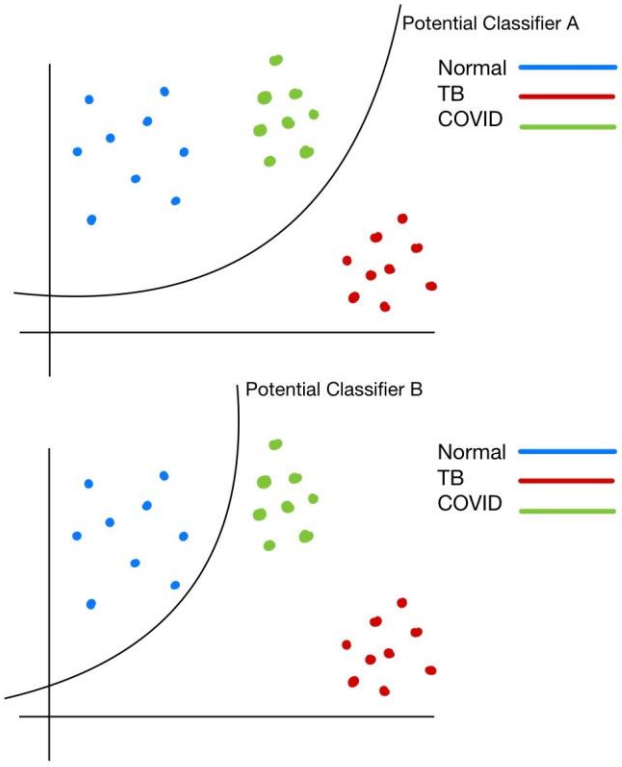

Figure 27: Potential classifier example

Illustrating what was just stated, Inception V3's decision boundary would look more like potential classifier B, while AlexNet's boundary would look more like potential classifier A. Regardless, the out-of-distribution data is an indicator of dataset similarity with respect to a model's decision boundary.

V. Conclusion

We explored the topics of trustworthiness of AI applications in tuberculosis diagnosis. For our project, we investigated the robustness of two classification neural networks, AlexNet and Inception V3, on image degradations and out-of-distribution data. The conclusion we have drawn from the project is that current advancement in classification models needs to take into account the degree of image degradation commonly observed in hospital settings and be able to flag degraded or out-of-distribution data that could affect model performance.

We would like to acknowledge the limitations of our

research. Firstly, there are many cases for the artifacts of X-ray images, and we only did a few degradations to simulate them. Secondly, due to a lack of data and time, we decided to randomly take 20 percent training data (3,500 Normal and 2,455 TB samples) as testing/degradation (700 Normal and 491 TB), a better way would be random split the training and testing/degradation datasets or try K-fold validation. Lastly, we didn't perform an ablation study to find the reasons that contributed to the bad performance of the two models of certain degradations, so we couldn't answer if the models will have a better performance if we use the degradation methods as augmentation in the training process. Future work could focus on more degradation for common artifacts (such as grid cutoff) and an ablation study. Our ultimate goal is to build a benchmark and augmentation data pipeline for X-ray AI models to help AI and BME engineers build the models and improve the robustness and trustworthiness of them.

Author Contributions

Desmond: TB dataset preprocessing, AlexNet model training, testing – 33.33%
Linsey: Inception V3 model training, testing – 33.33%
Clark: Image degradation methods – 33.33%

References

[1] "Studies from Grandview Medical Center Describe New Findings in Artificial Intelligence (2020 Acr Data Science Institute Artificial Intelligence Survey)." Women's health weekly (2021): 515–. Print.

[2] Kjelle, Elin, and Catherine Chilanga. "The Assessment of Image Quality and Diagnostic Value in X-Ray Images: a Survey on Radiographers' Reasons for Rejecting Images." Insights into imaging 13.1 (2022): 36–36. Web.

[3] Rahman, Tawsifur. "Tuberculosis (TB) Chest X-Ray Database." *Kaggle*, 14 June 2021, https://www.kaggle.com/datasets/tawsifurrahman/tuberculosis-tb-chest-xray-dataset.

[4] Viradiya, Preet. "Covid-19 Radiography Dataset." *Kaggle*, 22 May 2021, https://www.kaggle.com/datasets/preetviradiya/covid19-radiography-dataset.

[5] Rashmiranu. "Tuberculosis Chest X-Ray InceptionV3." *Kaggle*, Kaggle, 28 Apr. 2021, https://www.kaggle.com/code/rashmiranu/tuberculosis-chest-x-ray-inceptionv3.

[6] "Tf.keras.preprocessing.image.imagedatagenerator." *TensorFlow*, https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator.

[7] Pujara, Abhijeet. "Concept of AlexNet: Convolutional Neural Network." *Medium*, Analytics Vidhya, 9 June 2021, https://medium.com/analytics-vidhya/concept-of-alexnet-convolutional-neural-network-6e73b4f9ee30#:~:text=the%20completion%20afterward.-,The%20Architecture%20of%20AlexNet,over%20the%201000%20class%20labels.

[8] T, Adith Narein. "Inception V3 Model Architecture." *OpenGenus IQ*, OpenGenus IQ, 8 Oct. 2021, https://iq.opengenus.org/inception-v3-model-architecture/.

[9] Krizhevsky, Alex. "ImageNet Classification with Deep Convolutional Neural Networks." NeurIPS Proceedings, 2021, https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[10] Ioffe, Sergey, and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." ArXiv.org, 2 Mar. 2015, https://arxiv.org/abs/1502.03167.

[11] "Advanced Guide to Inception V3." *Google*, Google, https://cloud.google.com/tpu/docs/inception-v3-advanced.

[12] American Journal of Roentgenology. 2012;198: 156-161. 10.2214/AJR.11.7237