Adversarial Attacks on an AI-powered Medical Imaging System and Facial Recognition Systems

Somya Sharma

Master of Engineering, Computer Engineering

University of Guelph

Guelph, Canada

somyas20@gmail.com

Abstract

With the increasing use of machine learning in various applications, the privacy and security of sensitive data have become a major concern. Adversarial attacks have emerged as a significant threat to the integrity of machine learning models, allowing attackers to manipulate the input data and deceive the models into making incorrect predictions. In this context, various adversarial defense techniques have been developed to improve the robustness of machine learning models against such attacks. These defenses include approaches such as input preprocessing, adversarial training, and post-processing techniques such as denoising.

This Report discusses the problem of privacy in adversarial attacks and the current state-of-the-art defenses. Specifically, we focus on the effectiveness of denoising techniques in defending against adversarial attacks. Our results demonstrate that denoising is a promising defense technique against adversarial attacks and can significantly improve the robustness of machine learning models while preserving the privacy of sensitive data.

Keywords: Fast Gradient Sign Method (FGSM), Non-local mean (NLM), Bilateral filter (BF), Guided filter (GF) , TensorFlow, MobileNetV3, Transfer learning, Adversarial attacks, Adversarial defences, Privacy.

Table of Contents

Introduction

1.1 Background, Motivation and Significance:

Privacy in machine learning models has become an important concern due to the widespread use of such models in various applications. Adversarial attacks have been shown to be a potential threat to the privacy of machine learning models, which can result in incorrect predictions and leak sensitive information. Therefore, developing effective defense techniques to protect machine learning models from adversarial attacks is crucial.

In this context, we present a technical report that focuses on the use of transfer learning via MobileNetV3 Large for classification tasks, such as facial recognition and medical imagery. We discuss the potential of adversarial attacks and their impact on the accuracy of the classification models. Furthermore, we explore the use of non-local means denoising as a defense technique to improve the robustness of the classification models against adversarial attacks.

The motivation behind this work is to develop a comprehensive understanding of the challenges associated with privacy in machine learning models and to investigate effective defense techniques against adversarial attacks. Additionally, we aim to demonstrate the practical implementation of the transfer learning approach, non-local means denoising, and Flask web application framework in the development of a user-friendly classification system.

Overall, the technical report presents a detailed analysis of the experimental results obtained from the classification of facial and medical images. we will investigate the impact of adversarial attacks on facial recognition systems and explore Fast gradient sign method (FGSM) technique for generating adversarial examples. We will also examine defenses against adversarial attacks on facial recognition systems and evaluate their effectiveness. Through experimental results and analysis, we aim to provide insights into the vulnerability of facial recognition systems to adversarial attacks and potential strategies for improving their robustness and security.

The report highlights the effectiveness of Non-Local Means (NLM) denoising as a defense technique against adversarial attacks and provides insights into the practical implementation of transfer learning and Flask web application framework for classification tasks.

1.2 Objectives:

1. To build an image classification model using a feature extraction transfer learning on MobileNetV3 (trained on ImageNet).
2. To investigate the vulnerability of the model to adversarial attacks using Fast Gradient Sign Method (FGSM).
3. To explore non-local means (NLM) and its combination with other denoising technique as an adversarial defense mechanism and evaluate its effectiveness in restoring the classification accuracy of the model on adversarial images.
4. To develop a user-friendly Flask web application that allows users to upload and classify images using the trained model, as well as visualize the impact of adversarial attacks and defense.

Literature Review

2.1 Overview of Adversarial attacks and defense techniques:

Adversarial attacks have emerged as a major challenge in the development and deployment of facial recognition systems. In this literature review, we will examine some of the key studies and research in this area, focusing on the techniques for generating adversarial examples, the impact of adversarial attacks on the accuracy of facial recognition systems, and the defenses against adversarial attacks.

One of the most common techniques for generating adversarial examples is the Fast Gradient Sign Method (FGSM), which involves adding a small amount of perturbation to the input image to produce an incorrect classification.

Several studies have shown that adversarial attacks can significantly reduce the accuracy of facial recognition systems. In a study by Sharif et al. (2016), it was shown that adding imperceptible perturbations to facial images can cause the recognition accuracy to drop from 97.5% to 0%. Other studies have shown that adversarial attacks can be effective even in real-world scenarios, where lighting conditions and other factors can make it more difficult to generate successful attacks.

To counter these attacks, various defense mechanisms have been proposed, including adversarial training, input preprocessing, and image denoising. Input preprocessing involves modifying the input data to make it more difficult to attack, such as adding noise or resizing the image. Image denoising involves removing the adversarial noise added to the image to restore it to its original form.

2.2 Transfer learning and MobileNetV3:

Transfer learning is a popular technique in deep learning, where a pre-trained model is used as a starting point for training a new model on a different task. One of the popular pre-trained models is MobileNetV3, which is a lightweight and efficient convolutional neural network architecture designed for mobile devices.

MobileNetV3 consists of several building blocks such as inverted residual blocks and squeeze-and-excitation modules. The inverted residual blocks help to reduce the number of parameters and computational cost while maintaining the accuracy of the model. The squeeze-and-excitation modules help to capture the relevant features by recalibrating the feature maps using channel-wise attention.

Transfer learning via MobileNetV3 can be used for various computer vision tasks such as image classification, object detection, and segmentation. The pre-trained weights of MobileNetV3 on ImageNet can be fine-tuned or used as a feature extractor on a new dataset, which helps to improve the accuracy and speed of training.

Methodology

This section describes the methodology followed in this study to accomplish the stated objectives. The methodology comprises of the following steps: dataset preparation, model training, adversarial attacks, defense mechanism, and evaluation metrics. TensorFlow has been used as the framework.

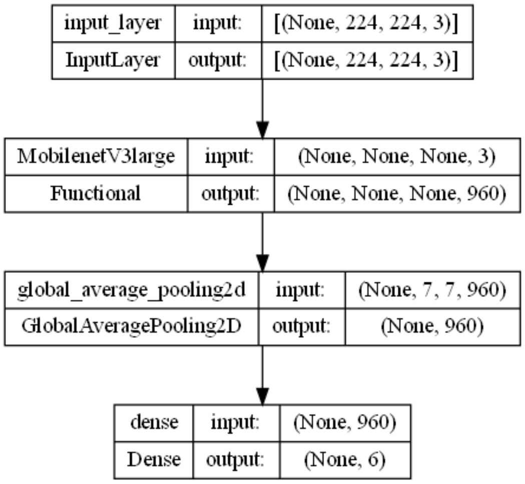
3.1 Dataset preparation and preprocessing:

We have used 5 Celebrity Faces Dataset and modified it to also include pictures of Somya Sharma. Training pictures include 105 Images. Validation pictures include 27 Images. Images are of 6 classes namely Ben Afflek, Elton John, Jerry Seinfeld, Madonna, Mindy Kaling and Somya Sharma. TensorFlow ‘s tf.keras.preprocessing.image_dataset_from_directory() function was used to import the dataset with a batch size of 1 and image size of 224x224 . Data was labeled and encoded as a categorical vector (for categorical_crossentropy loss). Similarly, we have used Chest X-ray (Covid-19 & Pneumonia) . Training pictures include 5144 Images. Validation/Testing pictures include 1288 Images. Images are of 3 classes namely Covid19, Normal, Pneumonia

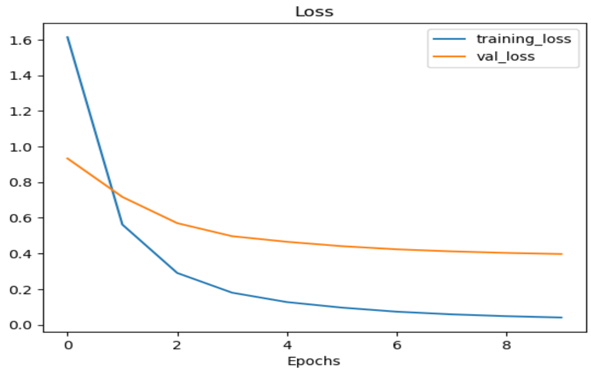
3.2 Model architecture and training:

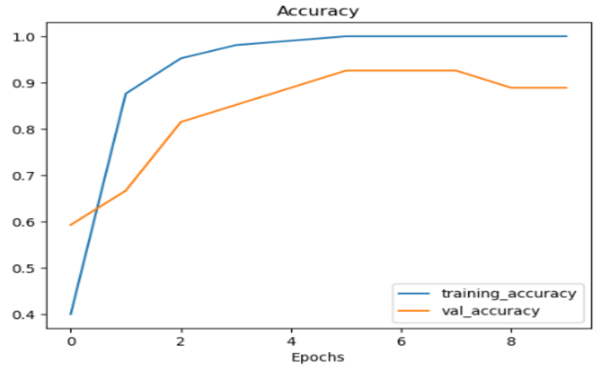
3.2.1 Celebrity Facial Dataset Model:

The model architecture used in this context is a type of machine learning model that employs transfer learning, which involves leveraging pre-trained models to extract features from input data. This feature extraction process is followed by a classification step, where the extracted features are used to classify the input data into one of six output classes, or categories. The input layer has a shape of (None, 224, 224, 3), which means the model expects input images with a size of 224x224 pixels and 3 color channels (RGB). The MobileNetV3 large model is used as a functional layer that takes the input tensor and returns an output tensor with shape (None, None, None, 960), where the first three dimensions are determined by the size of the input image and the fourth dimension corresponds to the number of filters in the last convolutional block of MobileNetV3 large. It uses MobileNetV3 large as the base model and has a global average pooling layer that reduces the spatial dimensions of the features and flattens them before passing them to a dense layer with 6 output units, one for each class. The model for the celebrity faces dataset was compiled using the categorical cross-entropy loss function, Adam optimizer, and accuracy as the evaluation metric. Default values of learning rate = 0.001 for Adam Optimizer was used for training the model.


input_layer | input: | [(None, 224, 224, 3)]
InputLayer | output: | [(None, 224, 224, 3)]
MobilenetV3large | input: | (None, None, None, 3)
Functional | output: | (None, None, None, 960)
global_average_pooling2d | input: | (None, 7, 7, 960)
GlobalAveragePooling2D | output: | (None, 960)
dense | input: | (None, 960)
Dense | output: | (None, 6)

Finally, accuracy is used as the evaluation metric to measure the performance of the model on the validation set during training. This metric calculates the percentage of correctly classified images out of all the images in the validation set. Model was trained for 10 epochs.

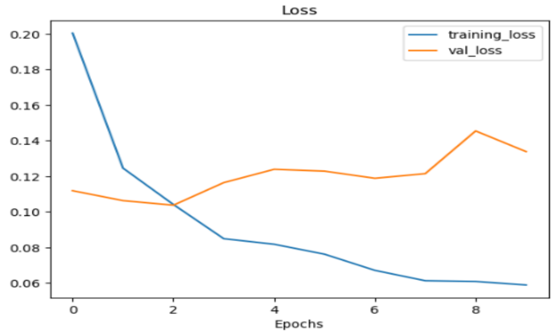


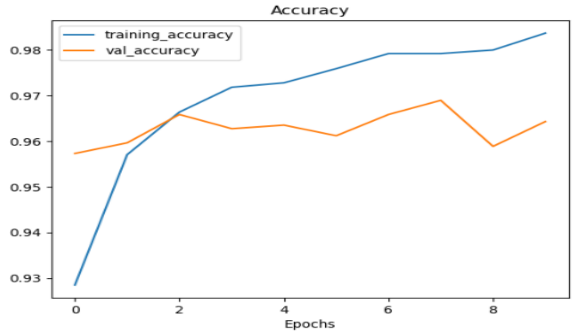


We achieved Training accuracy of 100% and Validation accuracy of 88.89% in Faces Model

3.2.2 Chest X-ray Dataset Model:

Model architecture is like Celebrity Facial Dataset Model except that it has 3 output classes.





We achieved Training accuracy of 98.37% and Validation accuracy of 96.43% in Chest X-Ray Model.

3.3 Adversarial attacks:

We will use the Fast Gradient Sign Method (FGSM). This technique has been shown to be effective in generating imperceptible perturbations to input images that can cause the facial recognition system to produce incorrect results. The code computes the adversarial perturbation for a given input image by calculating the gradient of the loss with respect to the input image, and then taking the sign of the gradient. This perturbation can be used to generate adversarial examples for testing the robustness of machine learning models.

3.4 Non-Local Means denoising technique:

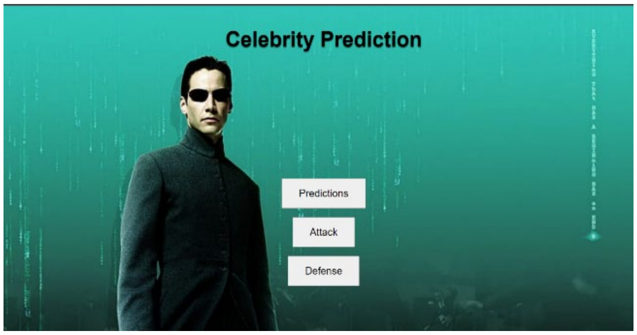
The Non-local Means (NLM) denoising technique is a widely used method for image denoising. It works by comparing each pixel in an image to its surrounding pixels and computes the average of the pixels with similar intensities. This helps to remove noise from the image while preserving the details and edges. The NLM technique has been shown to be effective in denoising a variety of images, including facial images, and has also been used as a defense mechanism against adversarial attacks. We have used OpenCV 's function cv2.fastNlMeansDenoisingColored() function for implementing NLM denoising.

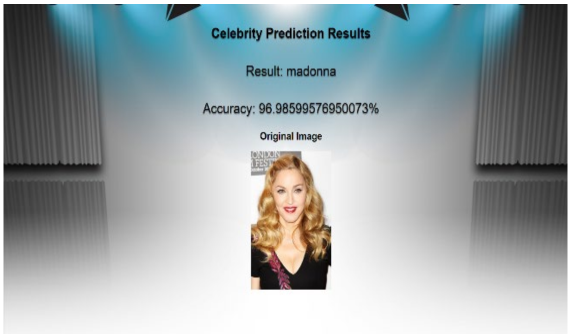
3.5 Bilateral filter and Guided filter:

The bilateral filter is a spatial-domain filter that smooths the image while preserving edges. OpenCV 's cv2.bilateralFilter() function has been used to implement bilateral filtering. The guided filter, on the other hand, is a linear filter that also preserves edges. OpenCV’s open cv 's cv2.ximgproc.guidedFilter() function has been used to implement for guided filtering.

3.6 Flask WebApp Implementation:

We implemented a simple user interface using HTML/CSS as frontend to provide a visually appealing and easy-to-use interface for the web application. To implement the Flask backend web application, we first created a folder structure containing the necessary files and folders, including templates, static files, and Python files. We then created the Flask app instance and defined the routes for the three functionalities: predicting the original image, predicting the adversarial attacked image, and predicting the adversarial defended image.


Celebrity Prediction

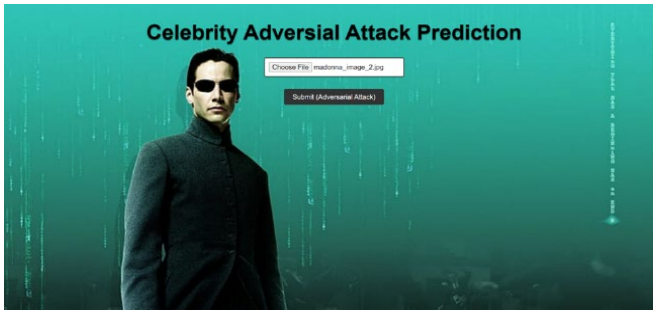

Celebrity Prediction Results

Result: madonna

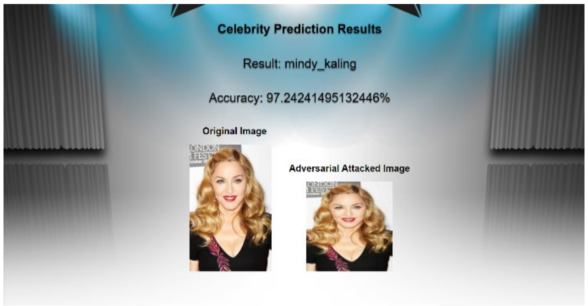
Accuracy: 96.98599576950073%

Original Image

To predict the original image, the user uploads an image, which is preprocessed and passed through the trained model. The predicted label and corresponding probability are displayed to the user.


Celebrity Adversial Attack Prediction

Choose File madonna_image_2.jpg


Celebrity Prediction Results

Result: mindy_kaling

Accuracy: 97.24241495132446%

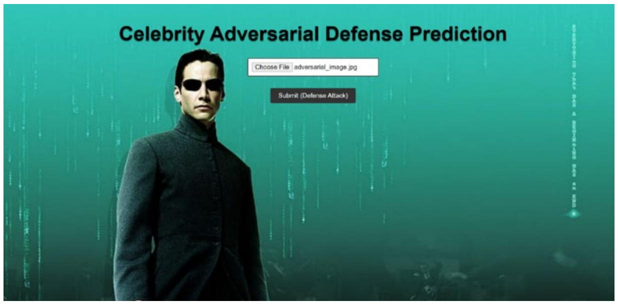
Original Image

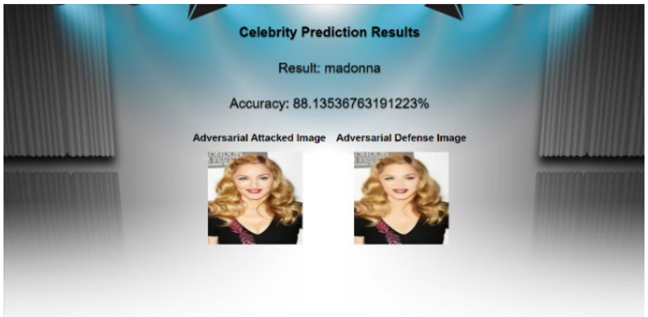
Adversarial Attacked Image

To predict the adversarial attacked image, we used the Fast Gradient Sign Method (FGSM) attack to generate a perturbed image. The perturbed image is passed through the same model, and the predicted label and probability are displayed to the user, along with the original image and the perturbed image. Perturbation value of 2 is used. The perturbed image is saved to *static/adversarial_image* .

To predict the adversarial defended image, we used the Non-Local Means denoising technique to remove the noise introduced by the FGSM attack. The user uploads the adversarial attacked image generated in previous step (present at *static/adversarial_image*) which is then denoised via NLM technique. The denoised image is then passed through the same model, and the predicted label and probability are displayed to the user, along with the perturbed (adversarial attacked image) image and the denoised image. The denoised image is also saved to *static/denoised_image* .

Two UIs:- Celebrity UI for Celebrity Facial Dataset and Chest X-Ray UI for Chest X-ray Dataset are provided as source code

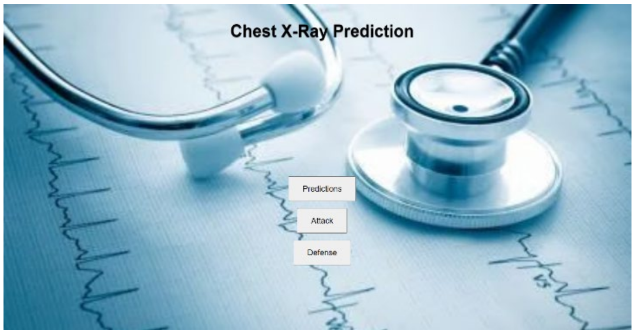

Celebrity Adversarial Defense Prediction

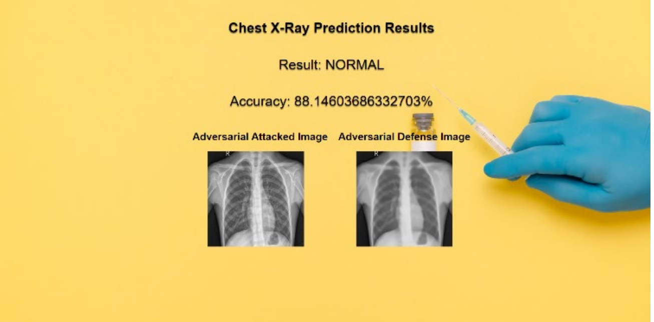

Celebrity Prediction Results

Result: madonna

Accuracy: 88.13536763191223%

Chest X-Ray UI :


Chest X-Ray Prediction


Chest X-Ray Prediction Results

Result: NORMAL

Accuracy: 88.14603686332703%

3.7 Jupyter Notebook Implementation:

We have also demonstrated many more examples, fully detailed inside the jupyter notebook included as source code (named: - adv attack Somya Faces.ipynb) for facial Dataset and (adv attack Chest_X_Ray Faces.ipynb) for Chest X-Ray Dataset.

Results and Discussion

In our experiments, we evaluated the effectiveness of using Non-Local Means (NLM) denoising technique as an adversarial defense method on a modified Celebrity Faces dataset. We also demonstrated the vulnerability of the model to adversarial attacks using Fast Gradient Sign Method (FGSM) and compared the results with and without NLM defense.

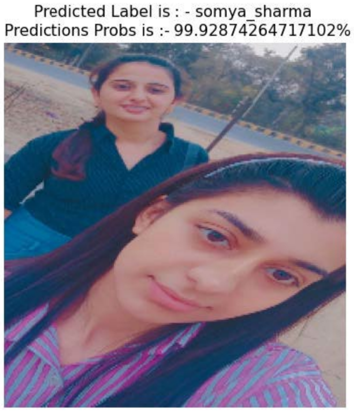
Our experiments showed that the model's accuracy decreased significantly when subjected to adversarial attacks using FGSM. However, when NLM denoising technique was applied to the adversarial attacked images, the accuracy was restored close to the original level. Which was generally further enhanced when it is combined with BF and GF denoising techniques

Furthermore, we implemented a Flask web application to demonstrate the predictions of the original images, adversarial attacked images, and adversarial defended images. The web application is an interactive tool for users to visualize the effect of adversarial attacks and the effectiveness of the NLM defense technique.

Overall, our results showed that NLM denoising technique can be an effective adversarial defense method to counter adversarial attacks on image classification models. NLM when combined with BF and GF techniques gives a even better result in most cases. Our web application can be a useful tool for researchers and practitioners to understand the impact of adversarial attacks and defense techniques on image classification models.

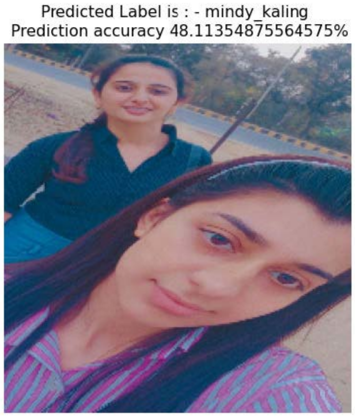
Here are two difficult examples, one with Somya's image along with someone and other of a celeb with a very high perturbation (=50)

Original Image: -


Predicted Label is : - somya_sharma
Predictions Probs is :- 99.92874264717102%

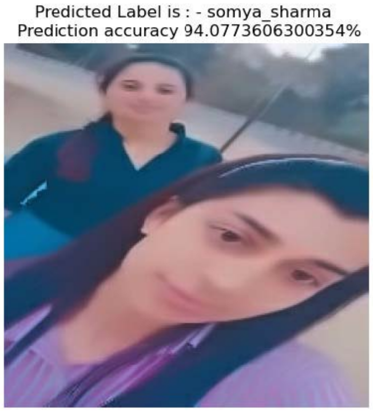
Perturbation Applied: -

Adversarial Attacked Image: -

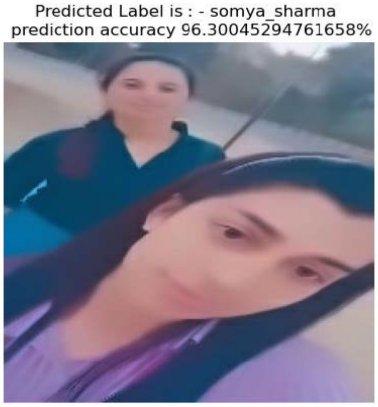

Predicted Label is : - mindy_kaling
Prediction accuracy 48.11354875564575%

Defended Images:

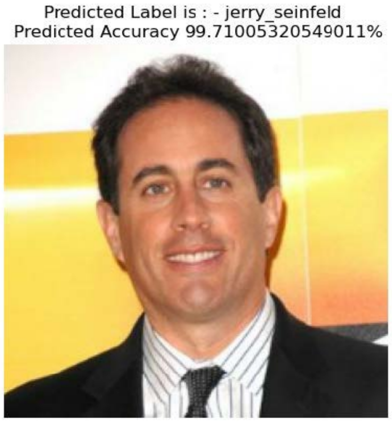
NLM


Predicted Label is : - somya_sharma
Prediction accuracy 94.0773606300354%

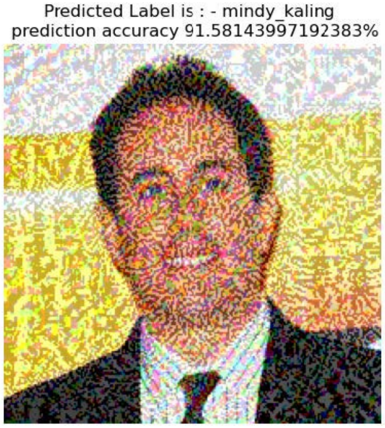
NLM + BF + GF


Predicted Label is : - somya_sharma
prediction accuracy 96.30045294761658%

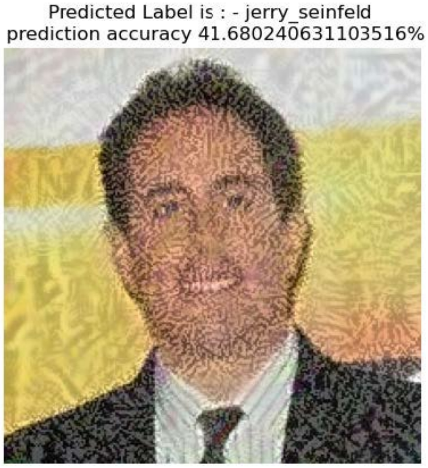
Original Image :- Perturbation Applied:- Adversarial Attacked Image:-

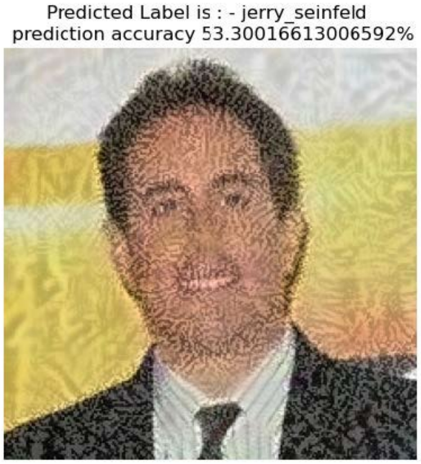

Predicted Label is : - jerry_seinfeld
Predicted Accuracy 99.71005320549011%

Perturbation Jerry


Predicted Label is : - mindy_kaling
prediction accuracy 91.58143997192383%

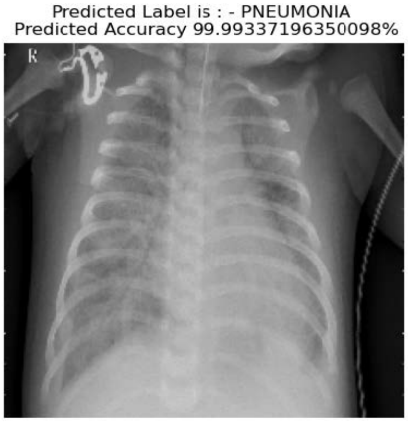
Defended Images: NLM NLM + BF + GF

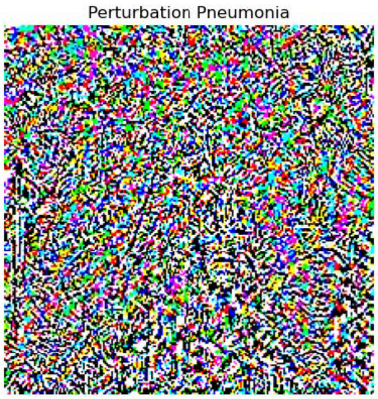

Predicted Label is : - jerry_seinfeld
prediction accuracy 41.680240631103516%

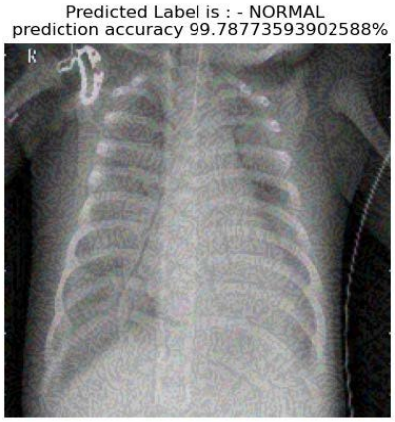

Predicted Label is : - jerry_seinfeld
prediction accuracy 53.30016613006592%

Now for Chest X-Ray Images, one picture of a pneumonia image with a high perturbation (=10)

Original Image: - Perturbation Applied: - Adversarial Attacked Image:

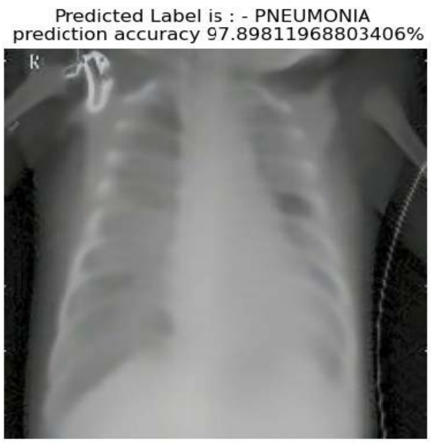

Predicted Label is : - PNEUMONIA
Predicted Accuracy 99.99337196350098%
R



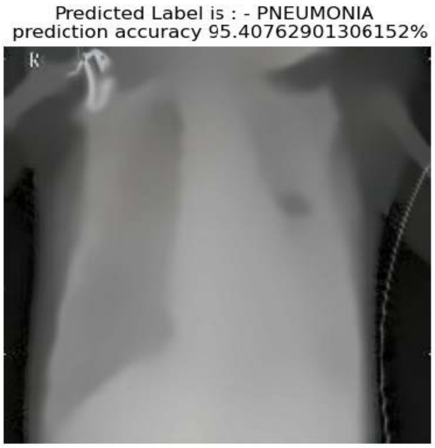

Predicted Label is : - NORMAL
prediction accuracy 99.78773593902588%
R

Defended Images:

NLM


Predicted Label is : - PNEUMONIA
prediction accuracy 97.89811968803406%

NLM + BF + GF


Predicted Label is : - PNEUMONIA
prediction accuracy 95.40762901306152%

Conclusion

5.1 Summary of the project:

This project aims to defend deep learning models against adversarial attacks by using non-local means denoising technique. It uses transfer learning and fine-tuning techniques on a modified Celebrity Faces and Chest X-ray Dataset to train a MobilenetV3 large model. The trained model is then subjected to adversarial attacks using FGSM, and the resulting adversarial images are defended using non-local means denoising technique. A flask web application is developed to demonstrate the effectiveness of the proposed defense technique.

Our project aimed to address the issue of privacy in AI by exploring adversarial attacks and defenses on facial and medical images. Adversarial attacks can compromise the privacy of individuals by manipulating images in a way that is not perceptible to humans but can deceive machine learning models. By applying the fast gradient sign method (FGSM) attack, we demonstrated how a facial image classification model can be easily deceived. We then showed how non-local means denoising, a well-known image filtering technique, can be used as an effective adversarial defense to mitigate the effects of the attack. We also combined our Non Local Means technique with other techniques like bilateral filters and guided filters which further increased the accuracy of the defended image.

In the context of medical images, we explored the use of adversarial attacks on chest X-rays to demonstrate their vulnerability to attacks. We then applied non-local means denoising as an adversarial defense and showed its effectiveness in restoring the original image. Our results show that adversarial attacks can be a serious threat to privacy in AI, and the use of appropriate defenses such as non-local means denoising can help protect against these attacks. We also combined our non-local means technique with other techniques like bilateral filters and guided filters which further increased the accuracy of the defended image. Overall, our project highlights the importance of considering privacy concerns when developing and deploying AI systems.

5.2 Future Work:

Exploration of other adversarial defense techniques such as adversarial training and feature squeezing. Testing the model on larger and more diverse datasets to further evaluate its robustness. Experimentation with different hyperparameters for the NLM, BF and GF denoising technique to achieve better results. Investigation of the use of other defense techniques like Adversarial Training, Defensive Distillation or combination of more Denoising techniques to further enhance the robustness of the model against adversarial attacks. Deployment of the web application to a cloud-based platform for easy accessibility and scalability.

References

1. M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2016.
2. J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," IEEE Transactions on Evolutionary Computation, vol. 23, no. 5, pp. 828-841, 2019.
3. H. Zhao, I. Gallo, I. Frosio, and J. Kautz, "Loss functions for neural networks for image processing," IEEE Transactions on Computational Imaging, vol. 3, no. 1, pp. 47-57, 2017.
4. S. Aneja, N. Aneja, P. E. Abas, and A. G. Naim, "Defense against adversarial attacks on deep convolutional neural networks through nonlocal denoising," ArXiv, abs/2206.12685, 2022.