TOWARD SECURITY AND PRIVACY ENHANCED DEEP NEURAL NETWORKS

by

DAVID RODRIGUEZ, M.S.

DISSERTATION
Presented to the Graduate Faculty of
The University of Texas at San Antonio
In Partial Fulfillment
Of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY IN ELECTRICAL ENGINEERING

COMMITTEE MEMBERS:
Ram Krishnan, Ph.D., Chair
Kal Clark, M.D.
Eugene John, Ph.D.
Ravi Sandhu, Ph.D.

THE UNIVERSITY OF TEXAS AT SAN ANTONIO
College of Engineering
Department of Electrical and Computer Engineering
August 2023

DEDICATION

I would like to dedicate this dissertation to my beautiful wife, children, mother, brothers, sisters, nephews and niece who have continually inspired me to work hard and stay motivated during all my academic studies.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor and committee chair Dr. Ram Krishnan for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me all the time of research and writing of this dissertation. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my dissertation committee: Dr. Ravi Sandhu, Dr. Kal Clark, Dr. Eugene John, for their encouragement, insightful comments, and hard questions.

Last but not least, I would like to thank my family: my wife and children for all the love and support during my academic studies.

This work is partially supported by NSF grants HRD-1736209 and CNS-1553696.

August 2023

TOWARD SECURITY AND PRIVACY ENHANCED DEEP NEURAL NETWORKS

David Rodriguez, Ph.D.
The University of Texas at San Antonio, 2023

Supervising Professor: Ram Krishnan, Ph.D.

Deep learning is transforming businesses with innovative technology in crucial industries such as manufacturing, transportation and healthcare. One example is medical imaging algorithms that are capable of diagnosing disease at a human expert level. Nevertheless, medical image deep learning models typically require large-scale image datasets and architectures to train state-of-the-art deep neural networks (DNNs). However, many raw image datasets contain sensitive identity feature information that prohibit entities from disclosing data due to privacy regulations. Additionally, large state-of-the-art DNNs are highly over-parameterized for medical image analysis. Consequently, medical image deep learning models are extremely vulnerable to adversarial attacks—imperceptibly perturbed input resulting in an incorrect model prediction. There are many security and privacy challenges that arise when developing DNNs for highly regulated industries such as healthcare. This work focuses on two major concerns that hinder the advancement of deep learning technology in crucial industries. First, data privacy during model development. Second, model robustness against adversarial attacks during model deployment.

This research develops learnable image transformation schemes. This topic is examined by investigating two image transformation schemes using convolutional autoencoder (CAE) latent representation and vision transformer (ViT) embeddings for privacy enhanced image classification. Additionally, this work includes an autoencoder-based image anonymization scheme that obfuscates visual image features while retaining useful attribute information required for model utility. The proposed anonymization method also enhances privacy by generating encoded images

that exclude sensitive identity feature information. Finally, this work develops an approach for adversarially robust deep learning model selection which includes an analysis on the role of deep learning model complexity in adversarial robustness for medical images.

TABLE OF CONTENTS

Acknowledgements . . . . . **iv**

Abstract . . . . . **vi**

List of Tables . . . . . **xv**

List of Figures . . . . . **xvii**

Chapter 1: INTRODUCTION & MOTIVATION **1**

1.1 Problem Statement 6

1.2 Summary of Contributions 7

1.3 Organization of Dissertation 7

Chapter 2: BACKGROUND & LITERATURE REVIEW **9**

2.1 Security and Privacy in Machine Learning 9

2.1.1 Secure Multi-party Computation 9

2.1.2 Homomorphic Encryption 10

2.1.3 Federated Learning 10

2.1.4 Visual Image Protection 10

2.1.5 Learnable Image Encryption 11

2.2 Adversarial Attacks on Deep Neural Networks 11

2.2.1 L-BFGS Attack 12

2.2.2 Fast Gradient Sign Method 12

2.2.3 Basic Iterative Method 12

2.2.4 One-Step Target Class Method 13

2.2.5 Projected Gradient Descent Method 13

2.3 Model Complexity and Sensitivity 13

2.3.1 Sensitivity and Generalization in Neural Networks 14

2.4 Defense Methods for Adversarial Attacks on Deep Neural Networks 15

2.5 Adversarially Robust Methods for Deep Neural Networks 15

2.5.1 Defensive Distillation 15

2.5.2 FGSM Adversarial Training 16

2.5.3 PGD Adversarial Training 17

2.5.4 Self-Supervised Learning for Adversarially Robust Networks 17

2.6 Medical DNNs in Adversarial Settings 18

2.7 Generating Adversarial Examples 20

2.7.1 Fast Gradient Sign Method 20

2.7.2 One-Step Target Class Methods 20

2.7.3 Basic Iterative Class Method 21

2.7.4 Iterative Least Likely Class Method 21

2.7.5 Project Gradient Descent Method 21

2.8 Threat Model 22

2.8.1 White-Box 22

2.8.2 Gray-Box 22

2.8.3 Black-Box 22

Chapter 3: Learnable Image Transformations 24

3.1 Evaluating Robustness of CAE and ViT Image Encoding for Privacy Enhanced Image Classification 26

3.1.1 CAE and ViT Image Encoding Formulation 26

3.1.2 CAE and ViT Image Encoding Datasets 28

3.1.3 CAE Network Architecture 28

3.1.4 ViT Network Architecture 29

3.1.5 Encoded Image Classification Model Architecture 29

3.1.6 CAE Encoded Image Training Procedure 29

3.1.7 ViT Encoded Image Training Procedure 30

3.2 Autoencoder-Based Image Anonymization Scheme for Privacy Enhanced Deep Learning 31

3.2.1 Image Anonymization Formulation 33

3.2.2 Multi-output Classification Loss Function 34

3.2.3 Identity Loss 34

3.2.4 Attribute Loss 35

3.2.5 Multi-output Classification Objective 35

3.2.6 Image Anonymization Loss Function 35

3.2.7 Identity Suppression Loss 36

3.2.8 Attribute Preservation Loss 36

3.2.9 Image Anonymization Objective 37

3.2.10 Image Anonymization Datasets 37

3.2.11 Anonymization Network Architecture 38

3.2.12 Multi-output Classification Model Training Procedure 39

3.2.13 Anonymization Model Training Procedure 40

Chapter 4: Adversarially Robust Deep Learning Model Selection 41

4.1 The Role Of Deep Learning Model Complexity In Adversarial Robustness For Medical Images 42

4.2 Medical Datasets 45

4.3 Model Complexity Network Architectures 46

4.4 Standard Training Procedure 47

4.5 Generating Adversarial Examples 47

4.6 Standard Training Experimental Setup 48

Chapter 5: Evaluation Process 51

5.1 CAE and ViT Image Encoding Evaluation 51

5.1.1 Encoded Image Classification Model Performance 52

5.1.2 CAE Public/Query Encoder Attack 52

5.1.3 ViT Public/Query Encoder Attack 53

5.1.4 Minimal Data Subset Attack 53

5.1.5 CAE Minimal Data Subset Attack 54

5.1.6 ViT Minimal Data Subset Attack 54

5.1.7 Reconstruction Cycle GAN Attack 54

5.1.8 Reconstruction Cycle GAN Adversarial Loss 56

5.1.9 Reconstruction Cycle GAN Cycle Consistency Loss 57

5.1.10 Reconstruction Cycle GAN Attack Full Objective 58

5.2 Image Anonymization Evaluation 59

5.2.1 Evaluating the Privacy/Utility Trade-off 59

5.2.2 Image Anonymization Evaluation with Classifier Transfer Attack 60

5.2.3 Image Anonymization Evaluation with Encoding Transfer Attack 60

5.3 Adversarial Robustness Evaluation 64

5.3.1 Adversarial Robustness & Model Complexity 64

5.3.2 Adversarial Attack Evaluation 64

Chapter 6: Experimental Results 67

6.1 CAE and ViT Image Encoding Results 67

6.1.1 Encoded Image Classification Model Performance 67

6.1.2 Public/Query Encoder Attack Results 68

6.1.3 Minimal Data Subset Attack Results 69

6.1.4 Reconstruction Cycle GAN Attack Results 70

6.2 Image Anonymization Results 71

6.2.1 Classifier and Encoder Transfer Attack Results 71

6.3 Adversarial Robustness Results 72

6.3.1 Cifar10 Model Performance 75

6.3.2 Mnist Model Performance 78

6.4 Saliency Maps of Adversarial Examples 81

6.4.1 Medical Image Saliency Maps 82

6.4.2 Cifar10 Saliency Maps 82

6.4.3 Mnist Saliency Maps 83

6.5 Decision Boundary Visualizations 87

6.5.1 Visualization Procedure 87

6.5.2 Medical Data Decision Boundaries 88

6.5.3 Cifar10 Decision Boundaries 90

6.5.4 Mnist Decision Boundaries 90

Chapter 7: Conclusion 93

7.1 Robustness Against Reconstruction Attacks 93

7.2 Autoencoder-based Image Anonymization 93

7.3 Robustness Against Adversarial Attacks 94

Bibliography 96

Vita

LIST OF TABLES

6.1 Privacy enhanced image classification accuracy. Model utility is preserved for DNN and ViT classifiers trained using encoded datasets. 68

6.2 CAE image reconstruction attack SSIM results. SSIM scores near 1 indicate high quality image reconstruction whereas scores closer to 0 indicate poor quality imge reconstruction. 69

6.3 ViT image reconstruction attack SSIM results. SSIM scores near 1 indicate high quality image reconstruction whereas scores closer to 0 indicate poor quality imge reconstruction. 70

6.4 Image Classification Accuracy of identity and attribute classifier for CelebA and Cifar-100 datasets 71

6.5 Classifier and encoding transfer attack performance on CelebA and Cifar-100 datasets 72

6.6 Chest X-Ray Average Accuracy, $\epsilon = 1$. 75

6.7 Dermoscopy Average Accuracy, $\epsilon = 0.2$. 75

6.8 OCT Average Accuracy, $\epsilon = 2$. 75

6.9 Cifar2 Average Accuracy, $\epsilon = 2$ 78

6.10 Cifar4 Average Accuracy, $\epsilon = 2$ 78

6.11 Cifar10 Average Accuracy, $\epsilon = 1$ 78

6.12 Mnist2 Average Accuracy, $\epsilon = 10$ 81

6.13 Mnist4 Average Accuracy, $\epsilon = 10$ 81

6.14 Mnist10 Average Accuracy, $\epsilon = 10$ 81

LIST OF FIGURES

3.1 The data owner uploads raw image dataset to MLaaS provider for the purpose of developing a deep learning algorithm directly using raw images. Adversary wishes to extract identity features. 25

3.2 The data owner transforms private dataset X_A using a private encoding function and transmits the encoded data Z_A and corresponding label Y_A to a cloud service provider. The attacker attempts to reconstruct the X_A using only $\{Z_A, Y_A\}$. 27

3.3 Privacy enhanced image classification using CAE encoding scheme. First, the CAE network is pre-trained using $x_b \sim p_{data}(x_a)$. Then, the DNN latent space classifier is trained using the latent representation and corresponding class labels. 30

3.4 Privacy enhanced image classification using ViT encoding scheme (image was adapted from [22]). First, the ViT network is pre-trained using $x_b \sim p_{data}(x_a)$. Then, the ViT embedding classifier is trained using the projection layer embedding space and corresponding class labels. 31

3.5 Image anonymization overview. 32

3.6 Examples of anonymized images from Celeba dataset using the proposed scheme.The bottom row are the corresponding anonymized images of the top row. 33

(a) Original images 33

(b) Anonymized Images 33

3.7 Proposed anonymization model architecture. 39

5.1 Cycle GAN reconstruction attack diagram. Where Z_B is the adversaries encoded dataset and Z_A is the data owner encoded set. Generator G_A translates the adversaries encoded set into the data owners encoded set domain. Generator G_B translates the data owner's encoded set into the adversaries encoded set domain. Discriminator D_B distinguishes between true adversary encoded images and fake adversary encoded images. Discriminator D_A distinguishes between true data owner encoded images and fake data owner encoded images. Decoder P_B reconstructs the adversaries encoded images. 55

5.2 Classifier transfer attack diagram. Where X'_A is the data owner's encoded dataset and X_B, Y_B are the attackers raw image dataset and identity labels which follows the probability distribution of the data owner's original dataset. I_B is the attacker's identity classifier. The attacker trains I_B with X_B, Y_B and uses the classifier to predict the identity label of the data owner's encoded dataset. 62

5.3 Encoding transfer attack diagram. Where X_B is the attacker's dataset which follows the probability distribution of the data owner's original dataset. X'_B is the attacker's encoded dataset which consists of attribute features and identity features. The data owners identity classifier is used to predict the identity label of the attacker's encoded dataset to verify if X'_B captures the data owner's identity features. 63

5.4 Set of optimal models with sufficient complexity for generalization. The set of optimal models must achieve low train and test error. 65

5.5 Adversarial robustness for a given set of optimal models of sufficient complexity. Adversarial examples are generated and all optimal models are attacked with increasing perturbation magnitude until model performance degrades. 66

6.1 The average accuracy and standard deviation of adversarial attacks on medical images. For all medical datasets the models of reduced complexity exhibit greater adversarial robustness, this is especially true for the PGD attacks. All networks exhibit similar performance on unperturbed data. . . . 74

6.2 The average accuracy and standard deviation of adversarial attacks on the Cifar10 datasets. For all versions of the Cifar10 datasets the models of reduced complexity exhibit greater adversarial robustness, this is especially true for the PGD attacks. Note that there is a small tradeoff between between accuracy and robustness for cifar10 as the models of lowest complexity generate slightly lower performance on unperturbed data but offer greater robustness prior to $\epsilon = 1$ for PGD attack. 77

6.3 The average accuracy and standard deviation of adversarial attacks on Mnist datasets. The adversarial robustness of mnist2 and mnist10 are mainly constant across all models. The mnist4 dataset demonstrates greater robustness for the model of reduced complexity while achieving comparable performance on unperturbed data. 80

6.4 Mnist saliency maps for clean (column 1 & 3) and adversarial images (column 2 & 4) generated with CNN6 and Resnet50 respectively. 84

6.5 The cifar10 saliency maps for clean and adversarial images generated with CNN6 and Resnet50. Attention regions remain fairly consistent across networks on clean data. 85

6.6 Mnist saliency maps for clean (column 1 & 3) and adversarial images (column 2 & 4) generated with CNN6 and Resnet50, respectively. 86

6.7 Adversarial examples on the decision boundary. Column 1 and 2 are the decision boundary visualizations for CBR-LargeT models before and after attacks, respectively. Column 3 and 4 are the decision boundary visualizations for Resnet-50 models before and after attacks, respectively. 89

6.8 Adversarial examples on the decision boundary. Column 1 and 2 are the decision boundary visualizations for CNN6 models before and after attacks, respectively. Column 3 and 4 are the decision boundary visualizations for Resnet-50 models before and after attacks, respectively. 91

6.9 Adversarial examples on the decision boundary. Column 1 and 2 are the decision boundary visualizations for CNN6 models before and after attacks, respectively. Column 3 and 4 are the decision boundary visualizations for Resnet-50 models before and after attacks, respectively. 92

CHAPTER 1: INTRODUCTION & MOTIVATION

Deep learning has achieved state-of-the-art performance in a variety of image classification tasks from natural image classification [82] to medical image analysis [77]. In particular, deep learning has progressively enabled faster and more accurate disease detection through the utilization of deep neural networks (DNNs) [77], which are multiple layers of interconnected nodes that learn through forward and backward propagation [27]. The development of medical imaging disease detection algorithms typically require large-scale image datasets and large state-of-the-art DNN architectures.

However, many raw image datasets contain sensitive identity feature information that prohibit entities from disclosing data due to privacy regulations such as Health Insurance Portability and Accountability Act, better known as "HIPAA" protect sensitive data from being released to the public. As a result, many deep learning practitioners are obliged to develop DNNs using *limited datasets*—limited set of identifiable patient information that the HIPAA Privacy Rule permits covered entities to share with certain entities for research purposes, public health or health care operations [85]. However, deep learning algorithms require large amounts of training data to generalize well in practice.

Additionally, training large deep learning architectures with large-scale image datasets requires expensive computational resources that many businesses simply cannot afford. As a result, cloud-based services have become an extremely popular option for data owners to outsource large computationally expensive deep learning tasks due to flexibility and cost saving [5]. Cloud providers offer a full range of services including storage, servers, virtual desktops, full applications and development platforms. Many organizations have access to large amounts of data but very limited computational resources and storage which prevent them from performing feature extraction tasks locally. Therefore, a large amount of data owners have opted for cloud services to allocate

resources as needed for the given task at hand [111]. Typically, an entity will send its raw data such as images to a machine learning as a service (MLaaS) provider for the purpose of developing a deep learning algorithm directly using the raw images. However, image data may contain sensitive information that the data owner wishes keep private while preserving model utility.

There are several privacy risks that accompany the disclosure of raw image data containing sensitive information. Raw images consist of features that are useful for a specific classification task such as classifying facial attributes which may include if an individual is smiling or wearing glasses, etc. On the other hand, raw images may also include additional feature information that is not useful for the specific classification task such as gender or age which could be used to reveal the identity of an individual. For example, previous work [104] has shown that person identification can be accomplished with as little as a human ear, so given a dataset of raw human faces an attacker could gain access to a victims personal identity by simply possessing an image of the human ear. Furthermore, [71] demonstrated that DNNs could be trained to recover patient identity from chest X-ray data by identifying if two frontal chest X-ray images belong to the same individual even if they were taken years apart. Attackers could potentially leak patient information or analyze the identified images to gain access to additional sensitive information. Consequently, this work aims to increase the privacy and security of sensitive data by transforming the original images such that sensitive information is excluded from encoded versions while maintaining classification accuracy.

One of the biggest challenges in accessing remote computing resources is keeping users and their data safe while working remotely. Data loss is the destruction of important or private information which is a major problem for users that are accessing external resources, it can be caused by theft, human error, viruses, malware, or power failures. Another challenge is data leakage which is the unauthorized transmission of data from within an organization to an external destination. This can be caused by a malicious insider, physical exposure, electronic communication or accidental leakage. Account hijacking is another security challenge which happens when cyber-criminals obtain login information to gain access to sensitive information stored on external

resources. There are also insider threats and insecure APIs that external resource providers must secure.

Cloud based attacks on deep learning models are another realm of challenges such as data poisoning attacks where an adversary pollutes training data to control model behavior. Model extraction attacks where the adversary queries a model on the cloud and uses the prediction to steal the models functionality. Model inversion attacks where an adversary tries to learn the training data by querying a model on the cloud. Adversarial attacks where an adversary adds imperceptible perturbation to data for the purpose of fooling a deep learning model. There are many security and privacy risks associated with accessing and utilizing external or cloud based resources. The first part of this work focuses on data privacy concerns when developing deep learning models.

Several visual information protection methods have been proposed to preserve privacy of image data such as pixelation, blurring and P3 [65]. Visual information protection methods encrypt data such that visible feature information of an image is concealed while making sure that the transformed version remains useful for classification [96], [92], [90], [91], [15]. However, these methods are prone to reconstruction attacks and do not exclude identity feature information from the encoded version of the original image. This research develops learnable image transformation schemes which obfuscate visual image features while preserving information required for deep learning classification model development. Additionally, this research develops an anonymization scheme that not only transforms the image such that it is longer recognizable to humans but it also excludes specific sensitive feature information from the encoded data.

One of the major challenges in developing algorithms to anonymize sensitive image data is known as the trade-off between privacy and utility [79], [56], [110]. The goal is to anonymize image data such that an attacker could not learn any sensitive identity feature information while authorized users could perform useful statistics. Eliminating the entire dataset provides perfect privacy but this is not useful. On the other hand, publishing raw unaltered data is statistically useful but may be detrimental to the privacy of sensitive data. This work proposes to publish

transformed versions of the original data that maintain model utility by retaining useful attribute features that are beneficial for classification while increasing privacy by removing sensitive identity features from the data.

Additionally, [95] showed that DNNs were vulnerable to adversarial attacks—manipulation of input data with imperceptible perturbation resulting in an incorrect model prediction. This is commonly referred to as an adversarial example. The goal of such an attack is to deceive the model into generating an incorrect output for a given input data point. This discovery has exposed a major weakness in DNNs. Furthermore, large state-of-the-art DNNs are highly over-parameterized for medical image analysis. Consequently, medical image deep learning models are extremely vulnerable to adversarial attacks. As a result, the reliability of DNNs has raised uncertainty as to whether they are safe and reliable in the physical world, especially in the medical domain.

Recent works have began investigating the domain of adversarial attacks on DNN models trained with medical images, these works suggest that medical DNNs are easier to attack than networks trained with natural images such as cifar10 and ImageNet. [26] confirms that medical DNNs are vulnerable to adversarial attacks and explains the motivation behind attacking such networks. This finding has motivated others to explore the extent to which medical DNNs are vulnerable to adversarial attacks. Further analysis by [58] discovered that medical DNNs are more vulnerable to adversarial attacks than DNNs trained with natural images, that is, adversarial attacks can succeed more easily on medical images using less perturbation than natural images. [58] claimed that this level of vulnerability inherent in medical images could be potentially due to the biological textures in medical images that may lead to high gradient regions that are sensitive to small perturbations. The utilization of overparameterized state of the art networks in the training process may also contribute to a sharp loss landscape for medical images. [76] analyzed the performance between models trained with state of the art architectures designed for ImageNet and smaller, simpler convolutional architectures for medical images, they found that the latter perform comparably to standard ImageNet models. This indicates that ImageNet performance is not predictive of medical

image performance.

Medical DNN models are particularly vulnerable to adversarial attacks due to the usage of over-parameterized networks on simple classification tasks [58]. It is common practice for practitioners to employ large state-of-the-art networks that were originally designed for natural images such as cifar10 [51] and ImageNet [82] on various classification tasks without assessing the adversarial robustness of the model—measure of the degree to which a DNN model can withstand an attack on the integrity and reliability of the network. The same is true for deep learning models utilized in realistic clinical settings. This convention exponentially increases the level of vulnerability found in medical DNN models. According to [62], model capacity is a crucial component of adversarial robustness for natural images—images captured in a natural setting, but there is a lack of documentation on the evaluation of adversarial robustness for medical diagnostic models with respect to model complexity. One way to overcome this problem is to evaluate model performance as attack strength increases and complexity is modified.

The features learned from medical images consist of simple biological textures that do not require large complex networks for feature extraction. Smaller,simpler networks can achieve comparable performance to state-of-the-art architectures on unperturbed data while producing models with greater robustness. This is also the case for cifar10. The mnist dataset exhibits similar performance and robustness across several architectural complexities. The goal of this study is to analyze model and data complexity for robust design in an adversarial setting.

An attacker could produce an adversarial example to generate the incorrect classification of a disease. This is a major problem that could result in a misdiagnosis. For example, an attack deployed against a skin cancer detection system could generate a benign output even if the original lesion was indeed malignant. An attack on natural images would not likely cause significant damage as opposed to an attack on medical images which could potentially cause major harm. In addition, attackers could be motivated by monetary gain through insurance fraud. For example, an attacker could submit fraudulent claims for related procedure charges with perturbed medical

images.

Nevertheless, data privacy and adversarial robustness of DNN models remain ongoing challenges in the development and deployment of security and privacy enhanced deep learning. This goal of this research is to develop learnable image transformation schemes using using convolutional autoencoder (CAE) and vision transformer (ViT) [22] to enable privacy enhanced image classification. Additionally, this work introduces a learnable image transformation scheme that enables image anonymization by the removal of identity feature information while preserving attribute features that are useful for model utility. The next goal of this research is to develop an approach for adversarially robust deep learning model selection. The aim of this research is to design robust networks that will mitigate adversarial attacks on medical diagnostic models. This study evaluates the adversarial robustness of DNN models trained with unperturbed image data and examines the relationship to model complexity.

1.1 Problem Statement

Deep learning models typically require large-scale image datasets and architectures to train state-of-the- art DNNs. However, many raw image datasets contain sensitive identity feature information that prohibit entities from disclosing data due to privacy regulations. Additionally, large state-of-the-art DNNs are highly over-parameterized for image classification in many highly regulated industries such as healthcare e.g., medical image analysis. Consequently, medical image deep learning models are extremely vulnerable to adversarial attacks. The advancement of deep learning technology is significantly hindered in highly regulated industries such as banking, healthcare, and insurance due to data privacy concerns and adversarial attack susceptibility.

1.2 Summary of Contributions

The contributions of this work are:

- Developed learnable image transformation schemes using convolutional autoencoder and vision transformer.
- Evaluate the robustness of CAE latent representation and ViT embedding image transformation schemes for privacy enhanced image classification.
- Demonstrate CAE latent representation and ViT embedding encoding schemes are robust to reconstruction attacks while preserving model utility.
- Developed an autoencoder-based image anonymization method for privacy enhanced deep learning.
- Increase privacy of image identity feature information while maintaining model utility.
- Developed an approach for adversarially robust deep learning model selection.
- Consider a set of medical image DL models that exhibit similar performances for a given task. These models are trained in the usual manner but are not trained to defend against adversarial attacks. This work demonstrate that, among those models, simpler models of reduced complexity show a greater level of robustness against adversarial attacks than larger models that often tend to be used in medical applications.

1.3 Organization of Dissertation

The remainder of the manuscript is organized as follows: Chapter 2 discusses related works in data privacy, image encoding and adversarial attacks in deep neural networks. A literature review

is included that compares many relevant works in the fields that were worked on in this dissertation. Chapter 3 outlines the methodology including CAE and ViT image encoding schemes and deep learning model complexity for adversarial robustness. Chapter 4 describes the evaluation procedure with dataset and model details. Chapter 5 describes the evaluation results. Chapter 6 summarizes and concludes the manuscript.

CHAPTER 2: BACKGROUND & LITERATURE REVIEW

The following section provides an overview of security and privacy in machine learning and fundamental first-order adversarial attacks and defenses on Deep Neural Networks. This section presents notable security methods and attacks in machine learning, image encoding and works in disease diagnostics using convolutional neural networks (CNNs). In addition, this section reviews some of the pioneering works of adversarial attacks in the medical imaging diagnostic domain. Furthermore, this section explores the relationship between model complexity and robustness for medical images natural images.

2.1 Security and Privacy in Machine Learning

Privacy protection in machine learning typically address the privacy of a model's input, the privacy of the model, or the privacy of the model's output. Several privacy preserving techniques have been proposed in the literature, some of which utilize secure multi-party computation, homomorphic encryption, federated learning, visual image protection and learnable image encryption.

2.1.1 Secure Multi-party Computation

Secure multi-party computation is a set of cryptographic protocols that allow multiple parties to evaluate a function to perform computation over each parties private data such that only the result of the computation is released among participants while all other information is kept private [108]. Secure multi-party computation methods have been applied in machine learning among multiple parties by computing model parameters using gradient descent optimization without revealing any information beyond the computed outcome [14], [67], [101], [69]. The investigated methods do not require multiple parties to perform gradient descent individually but instead allows all users to

anonymize private data individually and share the transformed images.

2.1.2 Homomorphic Encryption

Homomorphic encryption is a type of encryption that allows multiple parties to perform computations on its encrypted data without having access to the original data. It provides strong privacy but is computationally expensive requiring significant overhead to train machine learning models [4], [10], [17], [29], [47], [68]. The investigated encoding schemes do not require expensive encryption operations or specialized primitives for the training process.

2.1.3 Federated Learning

Federated learning allows multiple parties to train a machine learning model without sharing data [55], [9], [112]. For example, in centralized federated learning a central server sends a model to multiple parties to train locally using their own data, then each participant sends it's own model update back to the central server to update the global model which is again sent to each party to obtain the optimal model without access to the local data by iterating through this process [49]. Essentially, federated learning builds protection into the model. Nevertheless, federated learning suffers from the privacy-utility trade-off [44]. The investigated encoding schemes enable entities to share encoded data which do not reveal sensitive feature information and maintain model accuracy.

2.1.4 Visual Image Protection

Visual image protection methods transform original images to unrecognizable versions of the image while maintaining the ability to perform useful statistics. A few examples of visual image protection methods are pixelation, blurring, P3 [65], InstaHide [36] and NueraCrypt [107] which aim at preserving privacy and utility—a model trained on an encoded dataset should be approximately

as accurate as a model trained on the original dataset [11], [81]. InstaHide mixes multiple images together with a linear pixel blend and randomly flips the pixel signs. NeuraCrypt encodes data instances through a neural network with random weights and adds position embeddings to keep track of image structure then shuffles the modified output in blocks of pixels. The investigated encoding schemes remove the unnecessary complexity of NeuraCrypt's positional embeddings and permutations while maintaining privacy and utility.

2.1.5 Learnable Image Encryption

Learnable image encryption methods encrypt images such that the encoded versions are useful for classification [96], [92], [90], [91], [15]. However, in some cases network adjustments are required to process learnable image encryptions such as blockwise adaptation [96]. The investigated methods do not require any special modifications to the network and exclude identity information from the obfuscated samples while maintaining usability for classification. The proposed autoencoder-based image anonymization scheme is most closely related to [63] which removes user identity information from mobile sensor data while training a network to classify user activities. In this study, an autoencoder-based deep learning model is developed using image attribute features while removing image identity features.

2.2 Adversarial Attacks on Deep Neural Networks

An adversarial attack on deep neural networks is performed by generating a perturbation that is combined with the original input data sample. This is commonly referred to as an adversarial example. It is accomplished by manipulating the model's input data with the intention of producing an incorrect output.

2.2.1 L-BFGS Attack

[95] discovered that deep neural networks were vulnerable to adversarial attacks. The attack method implemented in [95] was the L-BFGS attack which utilized a linesearch optimization technique. However, the attack was considered computationally expensive.

2.2.2 Fast Gradient Sign Method

As a result, [28] introduced the Fast Gradient Sign Method (FGSM) attack, which is an untargeted single step max norm constrained attack method. The FGSM attack method generates adversarial examples by calculating the sign of the gradient of the loss with respect to the input data and constrains the amount of perturbation that can be added to each pixel of the original image with a max norm. The goal is to apply an imperceptible perturbation to the image that will result in a misclassification of the DNN model.

2.2.3 Basic Iterative Method

Later, [52] extended this work to an untargeted multi-step attack called the Basic Iterative Method and a targeted multi-step attack called the Iterative Least Likely Method. Both versions of the attack iterate through the previous FGSM process multiple times using a step size which acts as an epsilon for each iteration to find the optimal perturbation instead of adding the perturbation generated after single iteration. The Iterative Least Likely Method is a targeted attack that utilizes the argmin function to generate the label with the lowest probability of being the correctly predicted label to generate an adversarial example.

2.2.4 One-Step Target Class Method

The original single-step untargeted FGSM attack was later extended to a targeted attack in [53], which utilized a random label that was not likely to be equal to the true class or the least likely label to generate adversarial examples.

2.2.5 Projected Gradient Descent Method

The Projected Gradient Decent Attack Method (PGD) introduced by [62] is an extension of the BIM method that includes random restart from within the L-Infinity ball to generate the optimal perturbation. Random restart refers to the random location from within the L-Infinity ball where the FGSM process is initiated for each iteration on the original data point. The L-Infinity ball is generated with the max norm constraining epsilon as the radius value around an individual data point, this forms a maximum perturbation limiting boundary around the data point, if the data point is already randomly located near the upper bound and the step size is greater than the distance to the boundary then the perturbation is clipped to the edge of the boundary.

2.3 Model Complexity and Sensitivity

Resnets [32] were proposed to solve the vanishing gradient problem for very deep networks by adding skip connections to outperform shallow models. Although, skip connections in Resnet-like neural networks allow easy generation of highly transferable adversarial examples [103]. In fact, stronger adversarial examples were crafted by using gradients more from skip the connections rather than the residual modules. The more skip connections in a network the more transferable the attack. [103] investigates how skip connections affect the adversarial strength of attacks crafted on the network. identify one such weakness about the skip connections used by many state-of-the- art DNNs. the success rate drops more drastically whenever using gradients from a residual module

instead of the skip connection. This implies that gradients from the skip connections are more vulnerable (high success rate). Although these techniques are effective, they (as well as white-box methods) all treat the entire network (either the target model or the surrogate model) as a single component while ignore its inner architectural characteristics. The question of whether or not the DNN architecture itself can expose more transferability of adversarial attacks is an unexplored problem. If more gradients from the skip connections were used for an attack then the attack would be stronger and the model would experience greater degradation in performance. Could it be possible that input data features could contribute/impact the amount of gradients that flow through skip connections. identify a surprising security weakness of skip connections. Use of skip con- nections allows easier generation of highly transferable adversarial examples. Specifically, in ResNet-like (with skip connections) neural networks, gradients can backpropagate through either skip connections or residual modules. We find that using more gradients from the skip connections rather than the residual mod- ules according to a decay factor, allows one to craft adversarial examples with high transferability. Our findings not only motivate new research into the architectural vul- nerability of DNNs, but also open up further challenges for the design of secure DNN architectures. While different layers of a neural network learn different "levels" of features, skip connections can help preserve low-level features and avoid performance degradation when adding more layers. dependent on the amount of skip connections and the gradient flow through skip connections for the attack.

2.3.1 Sensitivity and Generalization in Neural Networks

This paper investigated the relationship between complexity and generalization by utilizing two metrics of complexity that focus on model sensitivity to input perturbations [70]. These metrics are the Jacobian norm and the number of transition. The input-output Jacobian of a trained model measures the sensitivity of the network to input perturbations by calculating the partial derivative of each element in the output probability vector (model prediction) with respect to the each element

in the input data (each pixel in an image). The Frobenius norm of the input-output Jacobian was utilized to estimate the sensitivity of functions. The sensitivity metric was utilized to study the behavior of models on and off the training data manifold. The authors found that sensitivity to perturbations correlates with generalization, i.e. as sensitivity increases the generalization gap (the difference between the train and test accuracy on all train and test data) also increases. The number of transitions was obtained by capturing the number of linear regions that a network splits the input space into. The Jacobian norm was utilized to measure the sensitivity of the linear regions. They found that trained models were more robust to perturbations that were in the vicinity of the training data manifold whose distance was found by the Jacobian norm. The authors claim that large neural networks often produce greater generalization as opposed to classical measures which is at odds with Occamâs razor. They claim that the input and function (model) should both be considered when evaluating complexity. The authors argue that Occamâs razor does not apply to neural networks since the best generalization can be obtained by much larger models.

2.4 Defense Methods for Adversarial Attacks on Deep Neural Networks

2.5 Adversarially Robust Methods for Deep Neural Networks

There are several defense techniques that have been proposed in the literature that attempt to produce greater robustness against adversarial attacks. Defending against adversarial attacks is an ongoing research effort. Here we review some notable works in defense methods that are readily available today.

2.5.1 Defensive Distillation

Distillation is a training procedure that utilizes knowledge transferred from large DNNs to models of reduced complexity without sacrificing performance [6] [34]. Specifically, distillation utilizes

the output vector probabilities of a previously trained model as the new label set to train a second model with the same data samples as the initial network on a smaller architecture. The probability vector values are amplified with a distillation temperature parameter for both models. Distillation allows the model to learn additional information about each training sample by leveraging the relationship of each label's probability in the initial model's output probability vector.

Defensive distillation is a technique to defend against adversarial attacks by utilizing the previously discussed distillation procedure with the exception of training the second distilled model with the same architecture as the initial network [73]. Recall that the original distillation process trains a second model with an architecture of lower capacity to reduce the computational complexity, whereas defensive distillation is not attempting to reduce computational complexity but to improve the robustness of the network to adversarial attacks. The intuition behind this method is as follows, training a model with probability vectors helps to prevent the model from fitting to tightly around the data, it contributes to a better generalization around the training points [73]. Model parameters are updated based on the additional knowledge gained from the relationship between classes in the probability vector and ultimately this decreases the amplitude of adversarial gradients.

2.5.2 FGSM Adversarial Training

Adversarial training is another defense technique that is utilized to resist adversarial attacks against deep neural networks by including adversarial perturbations in the training process. There are various forms of adversarial training but the main goal of this type of defense mechanism is to expose the network to adversarial perturbations and produce a robust model that is resistant to adversarial examples. [28] introduced adversarial training using the FGSM method to produce adversarial examples and include them in the training process to expose the network to the adversarial examples. This method trains a model using an adversarial objective function based on the FGSM attack method, this objective function is derived by obtaining 0.5% of the loss on the clean data and

0.5% loss from the adversarial data. This method was one of the forerunners in defenses against adversarial attacks.

2.5.3 PGD Adversarial Training

The projected gradient descent variant of adversarial training optimizes the saddle point (min-max) problem to produce robustness against a wide range of attacks [62].This is accomplished by first solving the inner maximization problem which is to maximize the loss with an adversarial perturbation from a set of allowed perturbations, in other words this describes the problem of generating an optimal perturbation that will result in the largest error of the model. Subsequently, the outer minimization problem is solved by finding model parameters that minimize the loss of the model trained on perturbed samples. Intuitively, this allows the model to prepare for the worst case perturbation within the L-infinity ball and resist the attack. The L-infinity ball can be seen as a space surrounding an individual data point that is specified by the max norm constraining factor referred to as epsilon which describes the magnitude of the perturbation. The radius from the data point is normally chosen as epsilon, thus an adversarially trained model should be able to withstand an attack from any perturbation that is generated within the ball. Solving the min-max optimization problem provides a guarantee that if an attacker is able to produce an adversarial example then it will not successfully fool the model. Thus if the adversarial loss is low for all perturbations then this means that generating adversarial examples should not be possible [62].

2.5.4 Self-Supervised Learning for Adversarially Robust Networks

Self-supervised learning is a method that is used to train a model for a supervised learning task with unlabeled data. Self-supervision was implemented as an alternative approach to transfer learning in [3]. Specifically, a model was trained using self-supervised learning on a large unlabeled dataset. The pre-trained lower layers were utilized for another learning task. This initializes a second model

with the weights of the lower layers of the pre-trained self-supervised network, then replaces and fine tunes the top layers for the new learning task with much less labeled data. The benefit is that the second model is a able to train with a much smaller dataset. They claimed that self-supervised learning produced models with greater performance and robustness to adversarial attacks than transfer learning on lower amounts of labeled data. Adversarial training with the self-supervised model offered greater robustness.

2.6 Medical DNNs in Adversarial Settings

Recent works have began investigating the domain of adversarial attacks on DNN models trained with medical images, these works suggest that medical DNNs are easier to attack than networks trained with natural images such as cifar10 and ImageNet. [26] confirms that medical DNNs are vulnerable to adversarial attacks and explains the motivation behind attacking such networks. This finding has motivated others to explore the extent to which medical DNNs are vulnerable to adversarial attacks. Further analysis by [58] discovered that medical DNNs are more vulnerable to adversarial attacks than DNNs trained with natural images, that is, adversarial attacks can succeed more easily on medical images using less perturbation than natural images. [58] claimed that this level of vulnerability inherent in medical images could be potentially due to the biological textures in medical images that may lead to high gradient regions that are sensitive to small perturbations. The utilization of overparameterized state of the art networks in the training process may also contribute to a sharp loss landscape for medical images. [76] analyzed the performance between models trained with state of the art architectures designed for ImageNet and smaller, simpler convolutional architectures for medical images, they found that the latter perform comparably to standard ImageNet models. This indicates that ImageNet performance is not predictive of medical image performance.

Medical DNN models are particularly vulnerable to adversarial attacks due to the usage of over-parameterized networks on simple classification tasks [58]. It is common practice for prac-

titioners to employ large state-of-the-art networks that were originally designed for natural images such as cifar10 [51] and ImageNet [82] on various classification tasks without assessing the adversarial robustness of the model—measure of the degree to which a DNN model can withstand an attack on the integrity and reliability of the network. The same is true for deep learning models utilized in realistic clinical settings. This convention exponentially increases the level of vulnerability found in medical DNN models. According to [62], model capacity is a crucial component of adversarial robustness for natural images—images captured in a natural setting, but there is a lack of documentation on the evaluation of adversarial robustness for medical diagnostic models with respect to model complexity. One way to overcome this problem is to evaluate model performance as attack strength increases and complexity is modified.

The features learned from medical images consist of simple biological textures that do not require large complex networks for feature extraction. Smaller,simpler networks can achieve comparable performance to state-of-the-art architectures on unperturbed data while producing models with greater robustness. This is also the case for cifar10. The mnist dataset exhibits similar performance and robustness across several architectural complexities. The goal of this study is to analyze model and data complexity for robust design in an adversarial setting.

An attacker could produce an adversarial example to generate the incorrect classification of a disease. This is a major problem that could result in a misdiagnosis. For example, an attack deployed against a skin cancer detection system could generate a benign output even if the original lesion was indeed malignant. An attack on natural images would not likely cause significant damage as opposed to an attack on medical images which could potentially cause major harm. In addition, attackers could be motivated by monetary gain through insurance fraud. For example, an attacker could submit fraudulent claims for related procedure charges with perturbed medical images.

One of the goals of this research is to design robust networks that will mitigate adversarial attacks on medical diagnostic models. Consequently, this study evaluates the adversarial

robustness of DNN models trained with unperturbed image data and examines the relationship to model complexity.

2.7 Generating Adversarial Examples

This section discusses the adversarial attack methods utilized to generate adversarial examples. The attack methods include the Fast Gradient Sign Method, One-Step Target Class Method, Basic Iterative Method, Iterative Least Likely Class method and Projected Gradient Descent attack Method.

2.7.1 Fast Gradient Sign Method

The Fast Gradient Sign Method (FGSM) introduced by [28] is a fast and simple way to generate adversarial examples. It is a max norm constrained attack that solves for the perturbation that maximizes the cost function. The max norm constraining factor limits the amount of change to an input image from the original pixel values. This method is a single step attack which attempts to solve for the optimal perturbation in a single iteration of back propagation.

$$\mathbf{X}^{adv} = \mathbf{X} + \epsilon sign(\nabla_X J(\mathbf{X}, y_{true})) \tag{2.1}$$

2.7.2 One-Step Target Class Methods

The One-Step Target Class Methods are an extension of FGSM that maximize the probability of a specific target label not likely to be the true label for a given input sample. The goal is to solve for a perturbation that minimizes the cost function for the true label and the target label [53].

$$\mathbf{X}^{adv} = \mathbf{X} - \epsilon sign(\nabla_X J(\mathbf{X}, y_{target})) \tag{2.2}$$

2.7.3 Basic Iterative Class Method

The Basic Iterative Method (BIM) is an extension of the FGSM attack. It performs FGSM for multiple iterations utilizing a step size alpha (α) to constrain the maximum allowed perturbation for each iteration [52].

$$\mathbf{X}_0^{adv} = \mathbf{X}, \quad \mathbf{X}_{N+1}^{adv} = Clip_{\mathbf{X},\epsilon}(\mathbf{X}_N^{adv} + \alpha sign(\nabla_{\mathbf{X}} J(\mathbf{X}_N^{adv}, y_{true})) \tag{2.3}$$

2.7.4 Iterative Least Likely Class Method

The Iterative Least Likely Method utilizes the least likely predicted class of a trained network for a given data sample [52].

$$y_{LL} = \arg\min_y \{p(y|\mathbf{X})\} \tag{2.4}$$

2.7.5 Project Gradient Descent Method

Projected Gradient Descent Method (PGD) is one of the strongest first-order attack methods which is an extension of FGSM. It iteratively attempts to produce an optimal perturbation from a random point within an L-Infinity ball. An epsilon value is utilized as the radius from the original data sample to produce the L-Infinity ball [62]. The PGD attack method is the inner maximization portion of the saddle point optimization problem in [62].

$$\mathbf{X}^{t+1} = \prod_{\mathbf{X}+s}(\mathbf{X}^t + \alpha sign(\nabla_{\mathbf{X}} J(\theta, \mathbf{X}, y))) \tag{2.5}$$

2.8 Threat Model

Threat modeling is a very important concept since it will determine the approach that the attacker will take when generating adversarial examples. It establishes the attackers knowledge regarding the training data, architecture and model parameters.

2.8.1 White-Box

Here the attacker is assumed to have perfect knowledge, that is, the attacker knows everything about the targeted system. The attacker has full knowledge of the target model's architecture and parameters. The perfect knowledge threat model allows security practitioners to perform a worst-case evaluation of the deep learning model.

2.8.2 Gray-Box

The attacker has limited knowledge of the target model. For example, the attacker may not know the exact training data but they may know the type of data utilized at test time. The attacker may also know the type of learning algorithm and architecture type but they do not know the model parameters.

2.8.3 Black-Box

The attacker has zero knowledge of the training data, architecture or model parameters. In this case the attacker is still able approximate the decision boundary of the target model by querying

the model and receiving feedback regarding the model's output.

CHAPTER 3: LEARNABLE IMAGE TRANSFORMATIONS

The utilization of deep learning image classification models has risen in the past several years, especially in highly regulated industries such as healthcare. Although, the development of state-of-the-art DNNs often requires large-scale image datasets. Nevertheless, data privacy concerns significantly hinder the advancement of deep learning technology.

Typically, data owners will either train deep learning models locally which means that computing resources can be accessed without a network or remotely which are computing resources that can be accessed through a network. Developing DNNs for image classification requires expensive powerful high-end hardware, including GPUs (graphical processing units). As a result, many businesses are switching to remote cloud-based options for DNN training.

Cloud-based services have become an extremely popular option for data owners to outsource large computationally expensive deep learning tasks due to flexibility and cost saving [5]. Cloud providers offer a full range of services including storage, servers, virtual desktops, full applications and development platforms. Many organizations have access to large amounts of data but very limited computational resources and storage which prevent them from performing feature extraction tasks locally. Therefore, a large amount of data owners have opted for cloud services to allocate resources as needed for the given task at hand [111]. Typically, an entity will send its raw data such as images and corresponding labels to a machine learning as a service (MLaaS) provider for the purpose of developing a deep learning algorithm directly using the raw images.

However, there are many privacy risks involved with uploading raw image datasets to MLaaS providers. Raw images may include sensitive feature information that data owners wish to keep private. They are susceptible to a wide range of attacks such as identity theft, disease misdiagnoses and insurance fraud [104], [71], [59]. For example, in Figure 3.1 the data owner wishes to upload face images and attribute labels such as smiling, attractive and beard to an MLaaS service

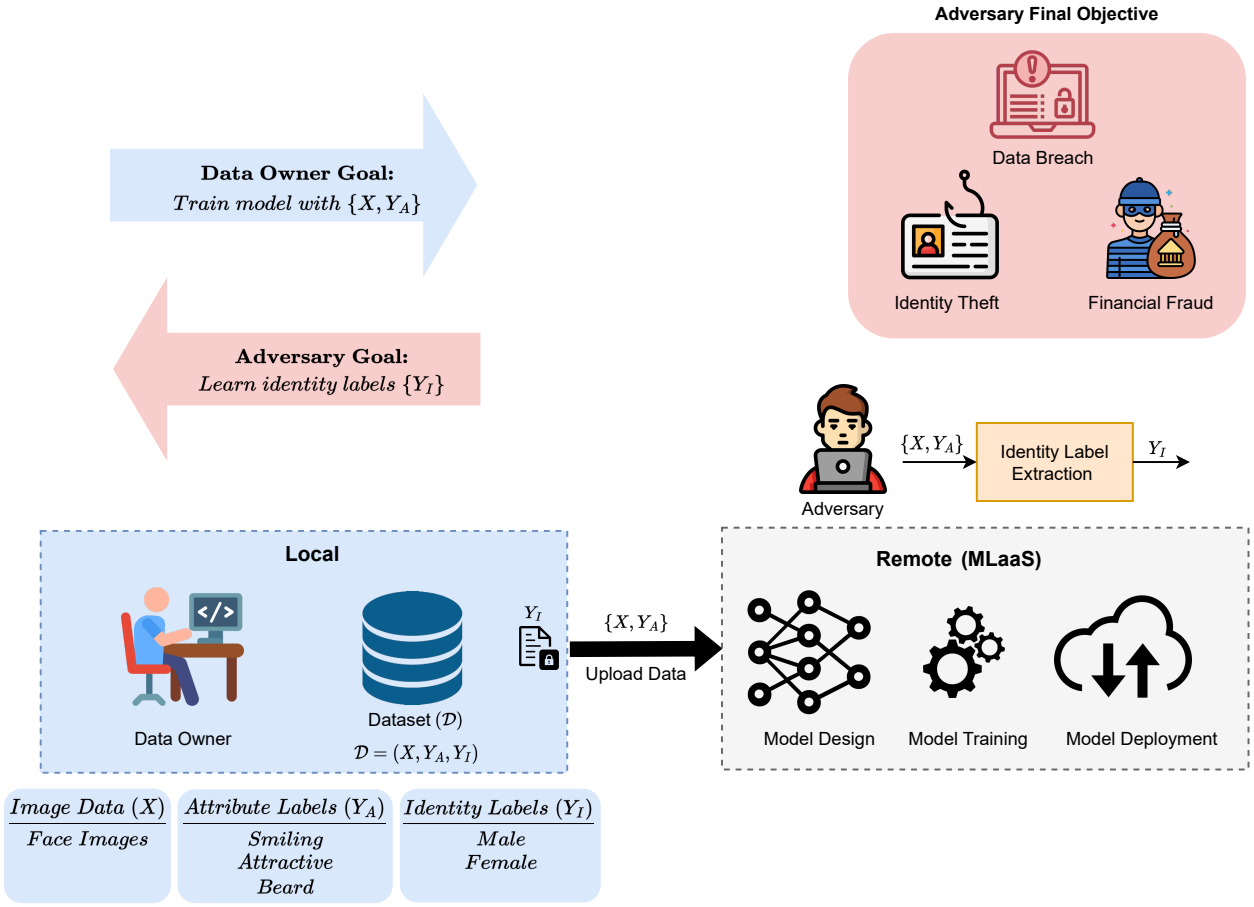


Figure 3.1: The data owner uploads raw image dataset to MLaaS provider for the purpose of developing a deep learning algorithm directly using raw images. Adversary wishes to extract identity features.

provider to train a DNN model while an adversary wishes to extract identity label information such as gender. The adversary could use raw face images to learn person identity features and gain access to private information.

Several visual information protection methods have been proposed to preserve privacy of image data such as pixelation, blurring and P3 [65]. Visual information protection methods encrypt data such that visible feature information of an image is concealed while making sure that the transformed version remains useful for classification [96], [92], [90], [91], [15]. Nevertheless, one of the major challenges in developing algorithms to encode sensitive image data is known as the trade-off between privacy and utility [79], [56], [110]. Additionally, these methods are

prone to reconstruction attacks and do not exclude identity feature information from the encoded version of the original image. This research develops learnable image transformation schemes that are robust to reconstruction attacks and exclude identity feature information. Learnable image transformations obfuscate visual image features while preserving information required for deep learning classification model development.

3.1 Evaluating Robustness of CAE and ViT Image Encoding for Privacy Enhanced Image Classification

* The material presented in this section is currently being reviewed to appear in the proceedings of the 10th European Conference On Service-Oriented And Cloud Computing (ESOCC 2023) in the article, "Evaluating Robustness of CAE and ViT Image Encoding for Privacy Enhanced Image Classification", co-authored with Ram Krishnan, Ph.D. and Yufei Huang, Ph.D.

This research investigates two image transformation schemes using CAE latent representation and ViT embeddings to enhance privacy of sensitive image data. The CAE latent representation is a compressed version of the original input data that captures important feature information relevant to image classification. Additionally, the ViT embedding consists of linear patch transformations with position embeddings. In both cases, the transformations are optimized to inherently reduce the feature space without any modifications to the network. Therefore, this work evaluates the robustness of CAE and ViT image transformation schemes against reconstruction attacks.

3.1.1 CAE and ViT Image Encoding Formulation

Let $\mathcal{X}$ be the set of all possible samples in the data domain, $X_a \subseteq \mathcal{X}$ is the data owner's private subset and Y_a is the corresponding label set. The data owner encodes $\{x_{a_i}\}_{i=1}^{N}$ where $x_{a_i} \in X_a$, using a private encoding function $z_a = E_a(x_a)$. The data owner's private samples are generated

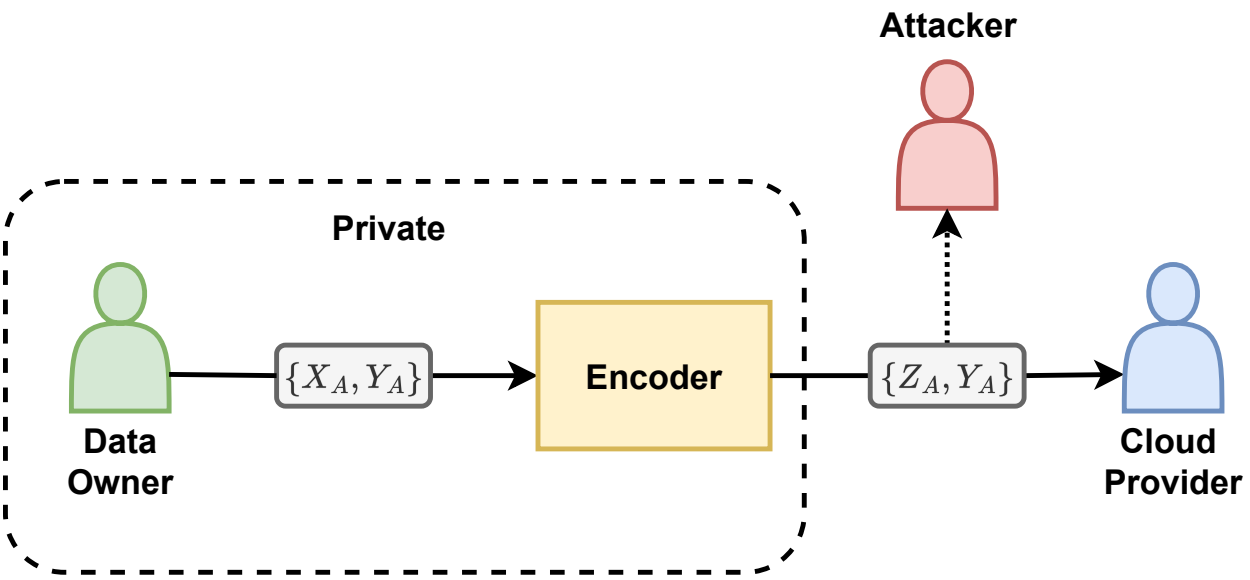


Figure 3.2: The data owner transforms private dataset X_A using a private encoding function and transmits the encoded data Z_A and corresponding label Y_A to a cloud service provider. The attacker attempts to reconstruct the X_A using only $\{Z_A, Y_A\}$.

according to the probability distribution $x_a \sim p_{data}(x_a)$. Next, the data owner shares the encoded set $\{z_{a_i}\}_{i=1}^{N}$ and corresponding class labels $\{y_{a_i}\}_{i=1}^{N}$ where $y_{a_i} \in Y_a$ with a third party cloud service provider to train a deeep learning classification model using the encoded samples without revealing sensitive data features. Afterward, the pre-trained network can be used to predict the class label given the encoded samples. This work aims to transform image data such that model utility is preserved while image reconstruction quality degrades. On the other hand, the attacker's objective is to learn a mapping function between the attacker's encoded set Z_b and the data owner's encoded set Z_a given that the attacker only has access to $\{Z_a, Y_a\}$ as depicted in Figure 3.2. This study assumes that an attacker is able to construct a dataset $X_b \subseteq \mathcal{X}$ and Y_b the corresponding label set that follows the probability distribution of the data owner's private subset $x_b \sim p_{data}(x_a)$. It is reasonable to assume that a dataset of similar probability distribution is available in practice e.g. suppose that the data domain is Chest X-ray images, then it is straightforward for an attacker to collect images from a publicly available Chest X-ray dataset. Also, this study assumes that the attacker has access to his own encoding function which is used to transform the constructed dataset, $z_b = E_b(x_b)$.

3.1.2 CAE and ViT Image Encoding Datasets

This research performs experiments using three publicly available datasets Chest X-ray [46], Fashion Mnist [21] and Cifar-10. The Chest X-ray dataset consists of 5,863 grayscale chest radiograph images of size 224x224 used to diagnose thorax disease. It includes two classes, where each image is labeled as "Pneumonia" or "Normal". The Fashion Mnist dataset consists of 60,000 train images and 10,000 test images. It includes grayscale fashion images of size 28x28 and associated label from 10 classes. The Cifar-10 dataset consists of 60,000 32x32 color images with 10 classes each having 6000 images. train images and 10,000 test images. It includes 50,000 train images and 10,000 test images.

3.1.3 CAE Network Architecture

The CAE encoder network used to encode Chest X-ray dataset consists of three convolution layers with 32, 64 and 128 filters, respectively. The model input size is 224×224. The CAE encoder network used to encode Fashion Mnist and Cifar-10 datasets consists of two convolution layers with 64 and 128 filters, respectively. The model input size is 28×28 and 32×32 for Fashion Mnist and Cifar-10, respectively. The kernel size is 3x3 with a stride of 2 and a latent space of 128. Each convolution layer consists of a leaky relu activation function with alpha 0.2 followed by a batch normalization layer. The decoder network used to reconstruct Chest X-ray datatset consists of three transposed convolution layers with 128, 64 and 32 filters, respectively. The decoder network used to reconstruct Fashion Mnist and Cifar-10 consists of two transposed convolution layers with 128 and 64 filters, respectively. The kernel size is 3x3 with a stride of 2 and output size of 224×224, 28×28 and 32×32 for Chest X-ray, Fashion Mnist, and Cifar-10 datasets, respectively. Each transposed convolution layer consists of a leaky relu activation function with alpha 0.2 followed by a batch normalization layer.

3.1.4 ViT Network Architecture

The ViT network used to encode all datasets consists of 8 transformer blocks and 6 attention heads. The hidden size is 1024 for Chest X-ray and 128 for Fashion Mnist and Cifar-10 datasets. The multi-layer perceptron (MLP) size is 2048 for Chest X-ray and 128 for Fashion Mnist and Cifar-10 datasets. There are 49 patches with size 32×32 for Chest X-ray. The Fashion Mnist and Cifar-10 datasets were resized to 48×48, thus there are 8 patches with size 6×6. In our evaluation, we use masked autoencoder (MAE) network [31] for ViT image reconstruction. MAE consists of a transformer based encoder and decoder network. We use a masking ratio of 75% with random sampling. The hidden size is 512 for Chest X-ray and 64 for Fashion Mnist and Cifar-10 datasets. The output size is 224×224 for Chest X-ray and 48×48 for Fashion Mnist and Cifar-10 datasets.

3.1.5 Encoded Image Classification Model Architecture

The CAE encoded image classification model consists of three fully connected layers (64, 16 and 2 hidden units) for Chest X-ray and (64, 16 and 10 hidden units) Fashion Mnist and Cifar-10 datasets. The ViT encoded image classification model follows the previously mentioned ViT network architecture for each respective dataset.

3.1.6 CAE Encoded Image Training Procedure

The CAE network is pre-trained using a dataset that follows the data owner's distribution. The mean squared error loss function is used to pre-train CAE with a batch size of 32 for 100 epochs. Afterward, the encoder network is used to transform image data into encoded samples i.e. the encoder latent representation. Next, a classification model is trained using the CAE encoded images as depicted in Figure 3.3. The network is trained using categorical crossentropy loss function and a batch size of 128 for 100 epochs. The network input is a 1D vector encoding of size 128. All

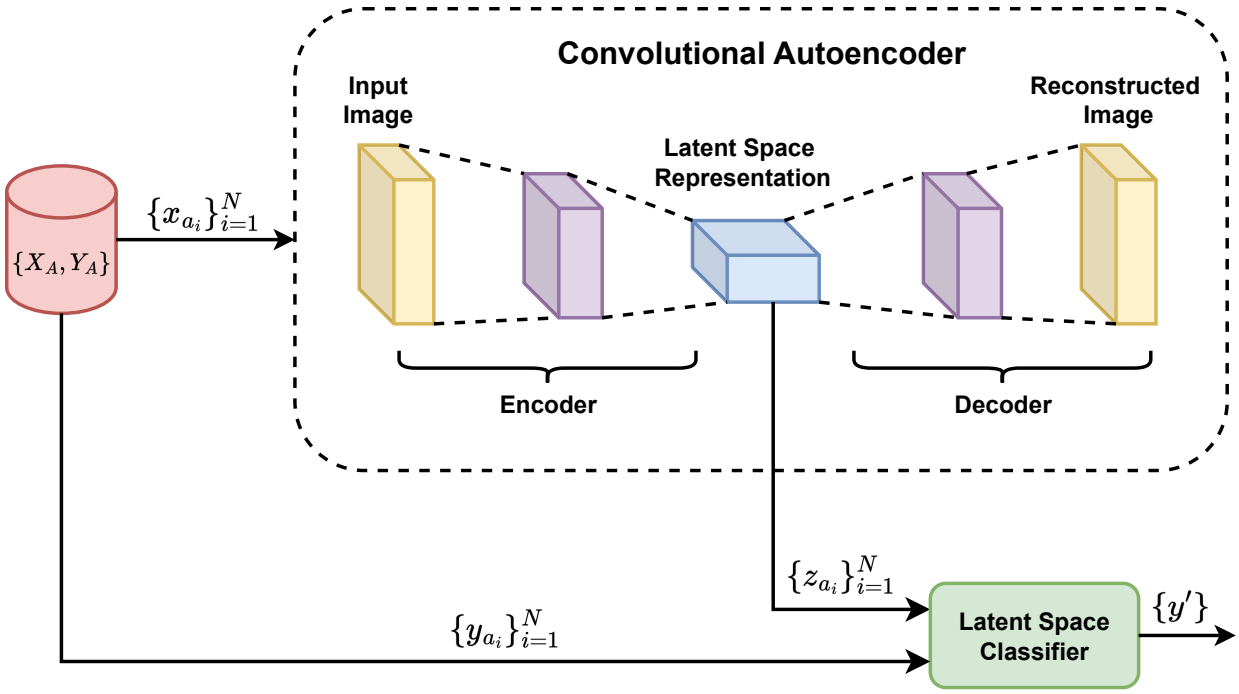


Figure 3.3: Privacy enhanced image classification using CAE encoding scheme. First, the CAE network is pre-trained using $x_b \sim p_{data}(x_a)$. Then, the DNN latent space classifier is trained using the latent representation and corresponding class labels.

training was completed using the Adam optimizer and tesla v100 graphical processing unit.

3.1.7 ViT Encoded Image Training Procedure

In the ViT experiments, the cross-entropy loss function is used to pre-train ViT with a batch size of 32 for 25 epochs. Afterward, the linear projection layer is used to generate patch embeddings which are added with position embeddings. Then, an embedding classifier is trained using a randomly initialized ViT network as depicted in Figure 3.4. All training was completed using the Adam optimizer and tesla v100 graphical processing unit.

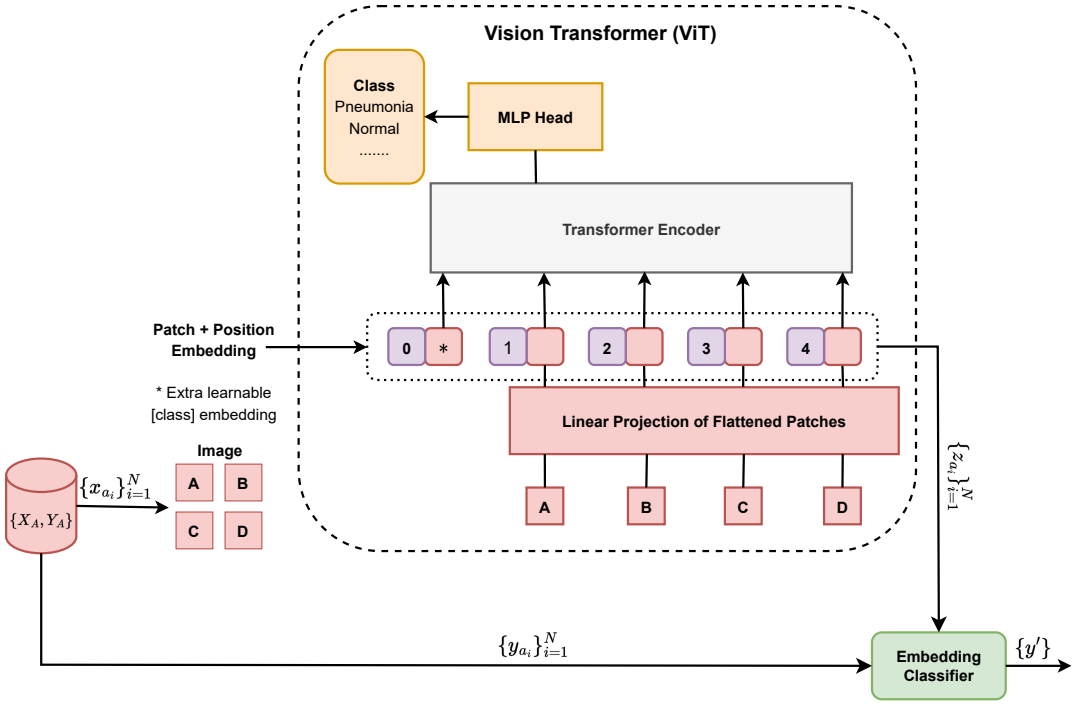


Figure 3.4: Privacy enhanced image classification using ViT encoding scheme (image was adapted from [22]). First, the ViT network is pre-trained using $x_b \sim p_{data}(x_a)$. Then, the ViT embedding classifier is trained using the projection layer embedding space and corresponding class labels.

3.2 Autoencoder-Based Image Anonymization Scheme for Privacy Enhanced Deep Learning

* The material presented in this section previously appeared in the proceedings of the 37th Annual IFIP WG 11.3 Conference on Data and Applications Security and Privacy (DBSec'23) in the article, "An Autoencoder-Based Image Anonymization Scheme for Privacy Enhanced Deep Learning", co-authored with Ram Krishnan, Ph.D.

Next, the goal of this research is to anonymize image data such that an attacker could not learn any sensitive identity feature information while authorized users could perform useful statistics. Eliminating the entire dataset provides perfect privacy but this is not useful. On the other hand, publishing raw unaltered data is statistically useful but may be detrimental to the privacy of sensitive data. This work proposes to publish transformed versions of the original data that

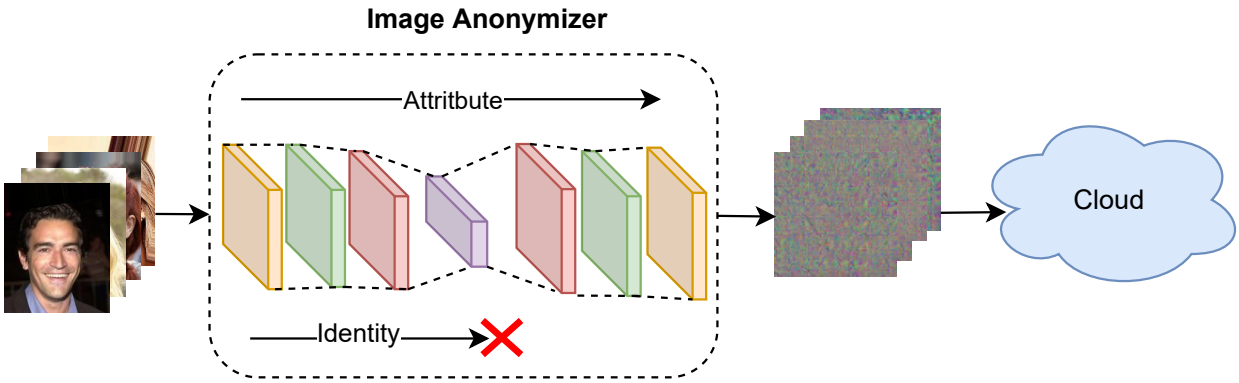
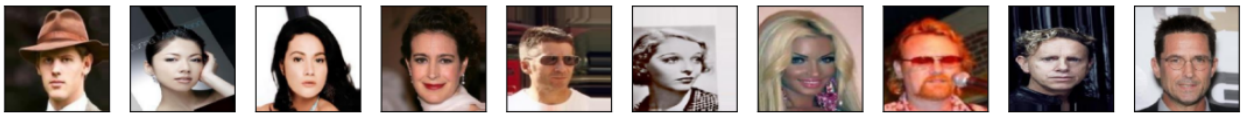


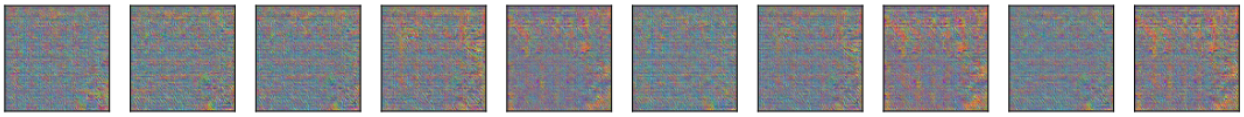
Figure 3.5: Image anonymization overview.

maintain model utility by retaining useful attribute features that are beneficial for classification while increasing privacy by removing sensitive identity features from the data. An overview of the image anonymization process is depicted in Figure 3.5. This research introduces an image data anonymization scheme using a deep learning approach to increase data privacy while maintaining model utility. Specifically, a multi-output deep learning model is trained to increase classification accuracy of identity feature information and image attributes. Then the anonymization network is trained which consists of a convolutional autoencoder attached to the input of a pre-trained multi-output classifier to generate obfuscated versions of the original images. The encoded images exclude identity feature information and preserve attribute features that are useful for classification.

The aim is to transform image data such that all visual feature information is unrecognizable to humans as depicted in Figure 3.6 but remains useful for classification. Additionally, the aim is to remove identity feature information from the transformed images while preserving attribute features. This method enables entities to share encoded versions of the original data that exclude sensitive feature information while maintaining model utility. This work considers identity features that can be collected from an image as sensitive data. On the other hand, this work considers attribute features in a given image as non-sensitive data. The objective is to preserve attribute features in the transformed images while removing sensitive identity features. Additionally, the objective is to maintain similar attribute classification performance on the transformed images as the original images.



(a) Original images



(b) Anonymized Images

Figure 3.6: Examples of anonymized images from Celeba dataset using the proposed scheme.The bottom row are the corresponding anonymized images of the top row.

3.2.1 Image Anonymization Formulation

Let $\mathcal{X}$ be the set of all possible 8-bit images in the data domain, $X_a \subseteq \mathcal{X}$ is the data owner's private subset and Y_a is the corresponding label set. Given the private image dataset $\{x_{a_i}\}_{i=1}^{N}$ where $x_{a_i} \in X_a$, the data owner encodes all images using a private image anonymization function $z_a = E(x_a)$ and shares the encoded set $\{z_{a_i}\}_{i=1}^{N}$ and corresponding attribute labels $\{y_{a_i}\}_{i=1}^{N}$ where $y_{a_i} \in Y_a$ with a third party cloud service provider without revealing sensitive identity feature information. The proposed image anonymization function is similar to [63] but instead of anonymizing mobile sensor data this work develops an encoding function to anonymize image data. The proposed method consists of a multi-output classifier to distinguish between attribute and identity features. In addition, the network consists of an autoencoder to anonymize images. The objective of training the network is to obtain the image anonymizer E^* which transforms raw images into anonymized images.

In the multi-output classification model training phase, a resnet50 model is trained to classify identity features and attribute features using the same input images $\{x_{a_i}\}_{i=1}^{N}$ with their respective class labels. The objective function for the multi-output classification model has two loss terms: identity loss for classifying identity features; and attribute loss for classifying attribute features. The aim is to classify identity features and attribute features of a given image with high classification accuracy for the multi-output network. After training, the multi-output classification

model is used to develop the anonymization network for the purpose of transforming original images into anonymized images. The anonymization objective function also contains two loss terms: identity suppression loss for removing identity features; and attribute preservation loss for preserving attribute features. The aim is to degrade the identity feature classification accuracy while preserving the attribute feature classification accuracy.

3.2.2 Multi-output Classification Loss Function

The multi-output network is trained using a multi-objective loss function for image classification which consists of an identity and attribute loss function. The identity loss is used to minimize the error between the true identity and identity classifier's predicted identity. The attribute loss is used to minimize the error between the true attribute and the attribute classifier's predicted attribute. The aim is to classify identity features and attribute features with high classification accuracy.

3.2.3 Identity Loss

The identity loss function L_i uses cross-entropy to measure the performance of identity classifier $I(\cdot)$ which is trained to classify image identity features.

$$L_i(I, X, Y) = -\frac{1}{N}\sum_{i=1}^{N} Y_i log(I(x_i)) \tag{3.1}$$

where x_i is the i^{th} image and Y_i is the corresponding ground truth identity label. $I(x_i)$ is the identity classifier's predicted output for the i^{th} image.

3.2.4 Attribute Loss

The attribute loss function L_a uses categorical cross-entropy to measure the performance of the attribute classifier $A(\cdot)$ which is trained to classify image attribute features.

$$L_a(A, T, X) = -\frac{1}{N}\sum_{i=1}^{N} T_i log(A(x_i)) \tag{3.2}$$

where T_i is the ground truth N-dimensional one hot encoded vector attribute label for the i^{th} image and $A(x_i)$ is the attribute classification function predicted softmax output which is an N-dimensional vector consisting of the attribute label probabilities for the i^{th} image.

3.2.5 Multi-output Classification Objective

The multi-output classification objective is:

$$L(I, A) = L_a(A, T, X) + L_i(I, X, Y) \tag{3.3}$$

The aim is to solve:

$$I^*, A^* =_{I,A} L(I, A) \tag{3.4}$$

3.2.6 Image Anonymization Loss Function

The image anonymization network is trained using a multi-objective loss function for image classification which consists of an identity suppression and attribute preservation loss function. The aim is to remove identity features while preserving attribute features that are useful for classification.

3.2.7 Identity Suppression Loss

The identity suppression loss function L_s uses mean squared error to remove identity feature information from sensitive data.

$$L_s(\xi, I^*, E) = -\frac{1}{N}\sum_{i=1}^{N}(\xi - I^*(E(x_i)))^2 \tag{3.5}$$

where E is the anonymization function and I^* is a pre-trained identity classification function. ξ is a positive value between 0-1. This work maximizes the difference between the predicted identity label and the true identity label by minimizing the mean squared error between ξ and the predicted identity label given the i^{th} encoded image. The anonymization network is penalized if the transformed image contains identity feature information.

3.2.8 Attribute Preservation Loss

The attribute preservation loss function L_p uses categorical cross-entropy to preserve attribute feature information.

$$L_p(A^*, E) = -\frac{1}{N}\sum_{i=1}^{N} T_i log(A^*(E(x_i))) \tag{3.6}$$

where A^* is the pre-trained attribute classification function. The aim is to minimize the preservation loss given the i^{th} encoded image. This work minimizes the difference between the predicted attribute label and the true attribute label by minimizing the crossentropy between T_i and the predicted attribute label.

3.2.9 Image Anonymization Objective

The image anonymization objective is:

$$L(\xi, I^*, A^*, E) = \lambda_1 L_p(A^*, E) + \lambda_2 L_s(\xi, I^*, E) \tag{3.7}$$

where the regularization parameters λ_1 and λ_2 are positive values that regulate the trade-off between privacy and utility.

The aim is to solve:

$$E^* = \arg\min_E L(\xi, I^*, A^*, E) \tag{3.8}$$

The anonymization function E^* generates encoded images that retain useful attribute features by penalizing the autoencoder network using crossentropy if the output does not contain attribute features. In addition, the autoencoder network is penalized using mean squared error if the output contains identity features. Thus the objective is used to preserve attribute features by applying L_p while removing identity features by applying L_s.

3.2.10 Image Anonymization Datasets

In this work, the publicly available CelebA [57] and Cifar-100 [**?**] image datasets are used to develop anonymization networks. The CelebA dataset is a large-scale face attribute dataset that consists of approximately 200K celebrity face images. It includes gender and 40 attributes per image with a variety of poses and backgrounds. However, in the experiments, this work selects images of 4 mutually exclusive attribute labels consisting of pale skin, smiling, eye glasses and wearing hat. Increasing the number of attributes significantly reduces the amount images per class.

Consequently, 10K images are included per attribute label. The goal is to train the anonymization network to generate encoded images that include attribute label features while removing gender label features. The Cifar-100 dataset consists of 60,000 32x32 color images. It consists of 100 classes containing 600 images each which are referred to as the fine label set. It is also available with 20 superclasses containing 3,000 images each which are referred to as the coarse label set. This work considers the fine label set to be private. Thus, the aim is to remove image features associated with the fine label set. The goal is to train the anonymization network to generate encoded images that include coarse label feature while removing fine label features.

3.2.11 Anonymization Network Architecture

The anonymization network architecture depicted in Figure 3.7 consists of two parts: a multi-output Resnet50 for image classification and a standard convolutional autoencoder (CAE) for image transformation. Resnets are large state-of-the-art DL architectures that consist of several blocks of residual modules and skip connections [33]. The multi-output architecture consists of one resnet50 feature extraction network with two separate classifiers at the output. The CAE encoder network consists of three convolution layers with 32, 64 and 128 filters, respectively. The kernel size is 3x3 with a stride of 2 and a latent space of 128. Each convolution layer consists of a leaky relu activation function with alpha 0.2 followed by a batch normalization layer. The decoder network consists of three transposed convolution layers with 128, 64 and 32 filters, respectively. The kernel size is 3x3 with a stride of 2 and output size of 224x224x3. Each transposed convolution layer consists of a leaky relu activation function with alpha 0.2 followed by a batch normalization layer.

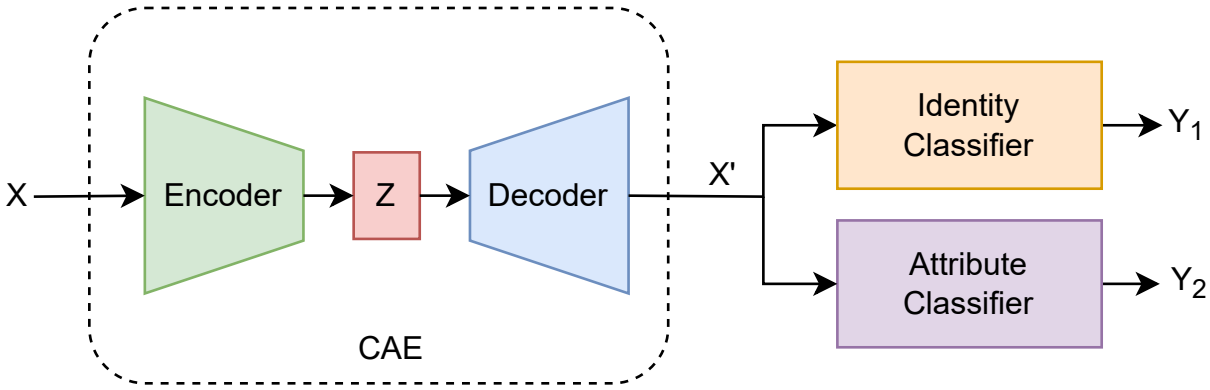


Figure 3.7: Proposed anonymization model architecture.

3.2.12 Multi-output Classification Model Training Procedure

The training procedure consists of a feature extraction phase for classification and an identity removal phase for anonymization. In the feature extraction phase the multi-output resnet50 model is trained from randomly initialized parameters for two different classification tasks given the same images. One classifier is trained to predict the gender identity for a given image using binary crossentropy loss function for the CelebA dataset. In the Cifar-100 experiments the identity classifier is trained using the fine label set which includes 100 classes. Simultaneously, a second classifier is trained to predict the attribute of the same image using categorical crossentropy loss function. In the Cifar-100 experiments the attribute classifier is trained using the coarse label set which includes 20 classes. The coarse label set is the superclass of the fine label set, e.g., the fish label is the superclass of aquarium fish, flatfish, ray, shark, trout. The aim is to classify fine label features and coarse label features for a given image set.

The network was trained using the adam optimizer with a batch size of 128. Check points were used to save the model with the highest validation accuracy during the training procedure. All images were resized to 224×224 and normalized between 0 and 1. The dataset was randomly shuffled and split to generate the train, test and validation set. Minor data augmentation was applied during training using keras image data generator which include zoom range 0.2 and horizontal flip. All training was completed using a tesla v100 graphical processing unit.

3.2.13 Anonymization Model Training Procedure

In the identity removal phase the CAE parameters are randomly initialized and its output is attached to the previously trained multi-output resnet50 classification model input. The resnet50 classifier model parameters frozen to ensure that the weights do not change during CAE training for the identity removal phase. During training the aim is to learn a CAE that retains useful attribute feature information to reconstruct an unrecognizable version of the original image for classification while removing the identity feature information. The identity classifier is optimized to remove identity feature information with a modified version of the mean squared error loss function. The attribute classifier is trained with the categorical crossentropy loss function to ensure that the anonymized images retain attribute feature information.

The network was trained using the adam optimizer with a batch size of 128. Check points were used to save the model with the highest validation accuracy during the training procedure. All images were resized to 224×224 and normalized between 0 and 1. The dataset was randomly shuffled and split to generate the train, test and validation set. Minor data augmentation was applied during training using keras image data generator which include zoom range 0.2 and horizontal flip. All training was completed using a tesla v100 graphical processing unit.

CHAPTER 4: ADVERSARIALLY ROBUST DEEP LEARNING MODEL SELECTION

The utilization of deep learning image classification models has risen in the past several years, especially in highly regulated industries such as healthcare. Although, the development of state-of-the-art DNNs often requires large complex deep learning architectures. Nevertheless, deep learning model complexity susceptibility to adversarial attacks significantly hinder the advancement of deep learning technology.

On the other hand, after the data is securely uploaded to an MLaaS provider for DNN model training, the next step is to design a network architecture for optimal performance. Typically, large state-of-the-art DNNs are used for training without considering the role that model complexity has on the network's vulnerability to attacks such as adversarial attacks which are discussed in detail in the literature review. Next, the model is trained using the dataset and architecture to obtain high classification performance. Finally, the trained network is deployed without assessing the model's susceptibility to adversarial attacks.

* The material presented next previously appeared in the proceedings of the International Conference on Intelligent Biology and Medicine (ICIBM 2021) in the article, "On the role of deep learning model complexity in adversarial robustness for medical images", co-authored with Tapsya Nayak, Ph.D., et al.

It is now well known that deep neural networks are vulnerable to adversarial attacks. As a result, many applications that depend on DNNs may be targeted by an attack which could have serious consequences, especially for safety critical industries such as healthcare. The healthcare domain has incorporated applications that utilize deep neural networks for the purpose of classifying disease but these models pose a security threat in that they are highly vulnerable to adversarial attacks. For example, an attacker might add imperceptible noise to a medical image in order to

cause a DNN-based medical diagnostic tool to misdiagnose disease. The lack of documentation on the adversarial robustness of medical DNN models has hindered the development and deployment of secure models in healthcare. One way to overcome this problem is to evaluate model complexity with respect to adversarial robustness. Medical images tend to focus on objects of interest consisting of various biological textures (spread of tiny features of various patterns) that do not require overparameterized networks for feature extraction. This work investigates how architecture size and complexity affects the robustness of DNN models trained on medical and natural images.

4.1 The Role Of Deep Learning Model Complexity In Adversarial Robustness For Medical Images

Next, this research evaluates the role of deep learning model complexity in adversarial robustness after training data has been securely uploaded to MLaaS providers for model development. Typically, large state-of-the-art DNNs are used for training without considering the role that model complexity has on the network's susceptibility to adversarial attacks. Instead, deep learning models are normally trained using large datasets and architectures to obtain high classification performance. After training, the network is deployed without assessing the model's susceptibility to adversarial attacks.

* The material presented in this section previously appeared in the proceedings of the International Conference on Intelligent Biology and Medicine (ICIBM 2021) in the article, "On the role of deep learning model complexity in adversarial robustness for medical images", co-authored with Tapsya Nayak, Ph.D., et al.

Deep learning has achieved state-of-the-art performance in a variety of image classification tasks from natural image classification [82] to medical image analysis [77]. However, deep learning models are vulnerable to adversarial attacks—imperceptible input perturbations utilized to produce an incorrect model prediction [95]. This inherent weakness in deep learning poses a

major security threat to medical deep learning models in that an attacker has the ability to alter the networks output. In fact, medicine may be uniquely susceptible to adversarial attacks [26].

Several defense techniques have been proposed to reduce model sensitivity to adversarial examples which include detection methods [106], defensive distillation [73], ensemble methods [97] and adversarial training [62]. Adversarial training is considered one of the most effective defense techniques. It minimizes the cost of a network trained on adversarial perturbations that maximize network error but suffers from performance degradation on unperturbed data [62]. Nevertheless, attaining adversarial robustness of deep neural networks remains an ongoing research effort.

Deep learning has been extensively utilized in the medical domain. Several deep learning based medical devices and algorithms in healthcare have been approved by the FDA to assist in diagnosing disease such as HealthPNX, Critical Care Suite & SubtleMR [8]. In fact, deep learning models have achieved remarkable performance for chest x-ray [77], dermoscopy [24] and retinal fundus classification [30]. However, medical image based deep learning models are also vulnerable to adversarial attacks [26]. Adversarial attacks against healthcare systems could interfere with proper medical diagnosis and potentially cause misdiagnosis by imperceptibly altering medical imaging that serve as input to DL based medical devices and algorithms in healthcare. These modifications may result in erroneous medical treatment and fraudulent billing to healthcare insurance providers [26]. Patient treatment plans can be changed by attacking Electronic Health Records (EHR), which is the digital version of patient medical records [2]. Attackers can produce adversarial examples to generate a specific disease prediction from medical image deep learning models. In fact, universal adversarial perturbations can achieve misdiagnosis at a very low cost and high success rate [35]. Furthermore, medical image deep learning models are more vulnerable to adversarial attacks than natural image DNNs, i.e., adversarial attacks can succeed more easily on medical images using less perturbation [58].

Generally, in the case of natural images, larger models are considered to be more robust

against adversarial attacks. In classical machine learning, the principle of Occam's Razor suggests choosing simpler models as they are expected to generalize better; however, larger ImageNet architectures often produce state-of-the-art performance in natural image classification [70]. As a result, Occam's Razor may not be a reliable heuristic for deep learning model selection in an adversarial setting. In fact, capacity is crucial for adversarial robustness [62], i.e., as capacity increases, natural image DL models become more resistant to adversarial attacks. Nevertheless, there is a trade-off between adversarial robustness and clean accuracy for natural image deep learning models [94]. However, the relationship between adversarial robustness and model complexity for medical image DL models has not been carefully studied.

Deep learning models deployed in realistic clinical settings often employ large deep learning architectures such as Resnet [32] for medical image classifications. However, these large Resnet trained on medical images do not significantly exhibit greater performance than smaller models [76]. Instead, smaller, simpler models provide comparable performance to large overly complex networks for unperturbed medical images. In fact, model complexity may have contributed to the high vulnerability of medical image deep learning models [58]. This was primarily attributed to a sharp loss landscape that was hypothesized to be the result of a highly complex network for a simple classification task. Instead, this work provides evidence that shows how model complexity influences adversarial robustness through decision boundary visualizations and saliency maps—image representation highlighting attention regions that influence a model's output the most [88]. A recent study [35] found that model architecture did not play a significant role in adversarial robustness for medical image deep learning models against universal adversarial perturbations. However, they only evaluate performance on state-of-the-art deep learning architectures, which are considered to be over-parameterized for medical image classification.

This work investigate whether simpler deep learning models of reduced complexity can produce comparable or improved robustness to state-of-the-art large networks for medical image classification. With this in mind, this research strives to understand *"How does model complexity*

impact adversarial robustness for medical image DL models"? *"Could models of reduced complexity offer greater robustness for medical image DL models"*?. To this end, this work investigates the role of model complexity in adversarial robustness for standard and adversarially trained medical image deep learning models.

Deep learning models are highly vulnerable to adversarial attacks for medical image classification. An adversary could modify the input data in imperceptible ways such that a model could be tricked to predict, say, an image that actually exhibits malignant tumor to a prediction that it is benign. However, adversarial robustness of deep learning models for medical images is not adequately studied. Deep learning in medicine is inundated with models of various complexity—particularly, very large models.

4.2 Medical Datasets

This research utilizes publicly available datasets to train all models. The medical datasets utilized in the experiments were Chest X-Ray, Dermoscopy and Optical Coherence Tomography (OCT). All models were trained with images of pneumonia class label or normal class label for the chest x-ray classification task. For the dermoscopy classification task all models were trained on images with label melanoma or not-melanoma, this work considers any non melanoma labeled image to be part of the not-melanoma class. The Optical Coherence Tomography (OCT) dataset was comprised of four classes consisting of CNV, DME, DRUSEN, NORMAL.

The chest xray dataset is publicly available on Kaggle [45], it consists of grayscale chest radiograph images used to diagnose thorax disease. It contains 5,863 chest radiographs with two classes.

The melanoma dataset is also publicly available on kaggle [83]. It contains 17.8K color images of skin lesions used to diagnose melanoma skin cancer. The data augmentation process from [80] was utilized for melanoma images.

The Optical Coherence Tomography dataset is publicly available at on kaggle [45], it consists of 84,495 grayscale images with four classes—choroidal neovascularization (CNV), drusen, diabetic macular edema (DME) or normal. Optical coherence tomography is a non-invasive imaging test that uses light waves to take cross-section pictures of the retina to assist in diagnosing retina disease and disorders in the optic nerve.

4.3 Model Complexity Network Architectures

The following experiments utilize a family of four Resnet architectures and a standard five layer Convolutional Neural Network (CNN) architecture for training all models. For experiments using the Resnet architecture, all models were initially trained with a publicly available Resnet50 architecture [16], which is a standard ImageNet architecture designed for images of natural scenery. Subsequently, the amount of layers were gradually reduced to produce three additional architectures which include Resnet32, Resnet20 and Resnet8. The CBR-LargeT architecture in [76] was utilized to produce the standard five layer CNN architecture. It consists of five convolutional layers, initially each layer is comprised of 32 filters and a 7x7 kernel size. The amount of filters are doubled at each convolutional layer while the kernel size remains constant for all layers. All convolutional layers were followed by batch normalization, relu activation and a max pooling layer with 3x3 window and 2x2 stride.

The complexity of a model can be altered through various methods such as increasing or decreasing the amount of layers, filters or kernel size in a network. The following experiments specifically reduce the amount of layers within the Resnet architecture, which was followed by a reduction in layer, filter and kernel size for the standard CNN architectures. [76] found that DNNs trained on medical images did not benefit much from utilizing large standard ImageNet architectures for training. Model capacity was reduced to assess adversarial robustness in an adversarial setting. The following experiments evaluate model complexity for adversarial robustness of medical and natural images.

4.4 Standard Training Procedure

All models were trained from random initialization of model parameters. [76] found that initiating the training process from pretrained ImageNet weights (transfer learning) did not significantly increase the performance of models trained with medical images. All models were trained using the Adam optimizer with a batch size of 32, learning rate scheduler and reduce learning rate on plateau. Checkpoints were utilized to store the model with the highest validation accuracy during the training procedure. All medical images were resized to 224x224 and all data was normalized between 0-1. Mnist was resized to 32X32 for compatibility with the resnet architecture. Each dataset was randomly shuffled and split using a random seed for reproducibility. The aforementioned training procedure was performed for a total of ten instances for each dataset and architecture to assess the average performance of all models across multiple subsets of the training data for a given architecture.

4.5 Generating Adversarial Examples

The attack methods deployed against the medical DNN models included the Fast Gradient Sign Method (FGSM) [28] and the Projected Gradient Descent (PGD) method [62] using the least likely class method for the target label. The magnitude of the perturbation was increased by utilizing a range of epsilon values where $\epsilon \in [0.01, 10]$ for each set of attacks. The ϵ value was extended up to 30 for the mnist dataset. FGSM is a single step max norm constrained attack that utilizes an epsilon value to restrict the amount of change or perturbation allowed in each pixel from the original pixel values in the input data. The PGD attacks utilized 20 iterations of back propagation updates to the input pixel values starting from a random initialization point within the L-Infinity ball to obtain the optimal perturbation with a step size of α = (ϵ * 0.1) for each attack and corresponding epsilon. The experiments were implemented in a targeted white-box attack setting as the source architecture and model parameters were known and utilized to generate adversarial examples.

4.6 Standard Training Experimental Setup

The experimental procedure consisted of training a DNN model with each of the previously specified architectures and datasets. Each dataset was combined, shuffled and randomly split to generate the train, test and validation set. The data selection process was repeated ten times for each architecture size. A single round of experiments consisted of training a model for each data split. This is equivalent to ten trained models for CBR-LargeT, resnet8, resnet20, resnet32 and resnet50, totaling 50 models for each dataset. For cifar10 and mnist datasets we replace CBR-LargeT with a standard 6 layer CNN architecture for compatibility reason. Adversarial examples were generated with a subset of the test data which the model was not previously exposed to during training and validation. Approximately, one hundred and fifty data samples were randomly selected from each class of the test set to generate adversarial examples. Data sample replacement was not utilized during the selection process.

Adversarial attacks were deployed on each model and the performance was assessed as model complexity was reduced by obtaining the average accuracy and standard deviation for all models trained on the same dataset and architecture. For example, the melanoma dataset was randomly shuffled and split ten times and for each data split we train a model using the resnet8 architecture which will result in ten trained models. This procedure was repeated for each dataset and architecture. The magnitude of the attack strength is limited by the max norm epsilon. As the epsilon value increases the amount of perturbation also increases which results in a higher amount of change to each pixel in an image. The model complexity of each network was reduced to model complexity as the attack strength was increased. The FGSM and PGD attacks were implemented using the Cleverhans library [72].

The medical datasets contain a small amount of classes so we investigate whether the number of classes contributed to our findings on medical DNNs. The robustness experiments were replicated on the cifar10 and mnist datasets while also reducing the number of classes down

equivalently to the medical datasets which have 2 classes for x-ray and melanoma and 4 classes for OCT dataset. It is typical for medical datasets to contain far less classes than natural image dataset and far less data samples.

Adversarial Training Procedure

Adversarial training was first introduced in 2015 [28], wherein they included adversarial examples into the training procedure to generate robust models. However, these trained models were still vulnerable as model robustness is directly related to the strength of adversarial samples being used during training. To address this in 2017, a new adversarial training algorithm that uses multi-step based PGD adversaries was proposed [62]. This achieves state-of-art robustness against L-infinity attacks on MNIST and CIFAR-10 dataset. A min-max formulation was used in training DL models [62]:

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} \mathcal{L}(\theta, x+\delta, y_{target}) \right] \tag{4.1}$$

where $\min_\theta \rho(\theta)$ represents the classification task, $\mathbb{E}_{(x,y_{target})}$ represents the empirical loss on the sample distribution $p_{x,y_{target}}$. The above saddle-point formulation is a composition of inner maximization and outer minimization problem. The former aims to find an adversarial version of x, using equation (4.1), to provide high adversarial loss, while the latter attempts to find model parameters θ to minimize the empirical classification loss. A previous study [62] found that robustness against PGD adversary provides robustness against all first-order adversaries and deep learning models with larger capacity can fit adversarial samples better. Motivated by the model performance using equation (4.1) on computer vision datasets, this study aims to evaluate the performance of medical deep learning models using equation (4.1) against adversarial and clean samples across different model capacities.

In this study, ResNet architectures of varying capacities - 8, 20, 32, 50 layers were trained

to generate adversarial trained models. The final layer for all the models were softmax with two neurons for Chest X-ray and Dermoscopy datasets, and four neurons for the OCT dataset. The networks were trained against adversarial perturbations that are max norm bounded. Each model was trained using initial weights from standard training of its counterpart network capacity, with learning rate of 0.001 and trained until the loss of the network would not further reduce or increase accuracy. To generate attacks during adversarial training, ϵ was set to 3/255, 1/255 and 10/255, with the step size set to $\epsilon/10$ and perturbation steps of 7, 5 & 5 for Chest X-ray, Dermoscopy & OCT datasets, respectively.

CHAPTER 5: EVALUATION PROCESS

This section describes the evaluation process of the CAE and ViT image encoding schemes for image reconstruction robustness. Also, included is the image anonymization evaluation for privacy enhanced image classification. Lastly, included is the evaluation on the role of deep learning model complexity in adversarial robustness for image data.

5.1 CAE and ViT Image Encoding Evaluation

* The material presented in this section is currently being reviewed to appear in the proceedings of the 10th European Conference On Service-Oriented And Cloud Computing (ESOCC 2023) in the article, "Evaluating Robustness of CAE and ViT Image Encoding for Privacy Enhanced Image Classification", co-authored with Ram Krishnan, Ph.D. and Yufei Huang, Ph.D.

First, this study evaluates the performance of the CAE and ViT encoded image classification models. Second, this study evaluates the robustness of CAE and ViT encoding schemes against four reconstruction attack methods which are refer to as *Public Encoder Attack, Query Encoder Attack, Image Subset Attack and Cycle GAN Attack*. In the experiments, two baseline reconstruction attacks are performed to assess the attackers ability to reconstruct original images given the assumption that the original encoding function is accessible. Also, this work assesses the attacker's ability to reconstruct original images given the assumption that a subset of original image-encoding pairs are available. Finally, this work assesses the attacker's ability to reconstruct original images given the assumption that only the encoded images and corresponding labels are available.

5.1.1 Encoded Image Classification Model Performance

This work evaluates the effectiveness of convolutional autoencoder (CAE) latent representation and vision transformer (ViT) [22] embeddings to preserve model utility using classification accuracy and robustness against reconstruction attacks using structural similarity index measure (SSIM).

First, we train the DNN latent space classifier and ViT embedding classifier using the data owner's encoded set $\{z_{a_i}\}_{i=1}^{N}$ and corresponding class labels $\{y_{a_i}\}_{i=1}^{N}$. Then, we evaluate the performance of each network using a test dataset of encoded images that were held out of the training process. The classification accuracy of the DNN latent space classifier shown in row 1 of Table 6.4 is 91.43%, 89.94% and 92.67% for Fasion Mnist, Cifar-10 and Chest X-ray datasets, respectively. The classification accuracy of the ViT embedding classifier shown in row 2 of Table 6.4 is 87.32%, 86.01% and 91.98% for Fasion Mnist, Cifar-10 and Chest X-ray datasets, respectively. We obtain high classification accuracy on encoded images using our DNN latent space classifier and ViT embedding classifier. This result demonstrates that model utility remains high using encoded datasets.

5.1.2 CAE Public/Query Encoder Attack

This study evaluates the robustness of CAE and ViT based image transformation schemes against attacks that aim to reconstruct original images given the strong assumption that an attacker has full access to the data owner's original encoding function. In the CAE experiments, the attacker concatenates a randomly initialized decoder model P_B to the data owner's original encoder. Then the attacker generates encoded data samples by using his constructed dataset as input to the data owner's original encoder. Finally, P_B is optimized by updating the model parameters based on the gradients of the mean squared error loss between the attacker's constructed dataset and the decoder's predicted output given the corresponding encoded samples i.e. $\hat{x}_b = P_B(E_A(x_b))$. During training, the data owner's original encoder weights are frozen. Afterwards, the decoder is used to

reconstruct the data owner's encoded dataset. Similarly, this work considers the scenario where an attacker only has access to query the data owner's original encoder network i.e., *Query Encoder Attack*. In this case, the CAE decoder is trained by minimizing the error between the queried output and the ground truth image.

5.1.3 ViT Public/Query Encoder Attack

In the ViT experiments, the attacker generates encoded samples using linear projection layer. Then, tokens are generated by adding position embeddings to the linear projection and shuffled to perform masking i.e. randomly remove tokens. The MAE encoder network is applied to the randomly selected visible tokens for training. The masked tokens are then concatenated to the MAE encoder output and unshuffled. Then, position embeddings are added and applied to the decoder's linear projection layer to train the decoder using all masked and visible patches. The goal of MAE is to reconstruct the masked patches in the pixel space. To do this, the mean squared error loss is computing between reconstructed and original images on the masked patches only. Afterwards, the data owner's ViT encoded images are applied to the MAE network for image reconstruction. Similarly, in the query encoder ViT experiments, the attacker generates encoded samples by querying the projection layer given the constructed dataset. Afterwards, the MAE network is trained using the previously mentioned procedure.

5.1.4 Minimal Data Subset Attack

Next, this study evaluates the robustness of CAE and ViT encoding schemes against minimal data subset attacks where the adversary gains access to a subset of the data owner's original dataset with the goal of training a deep learning model to reconstruct original images given encoded images. The attack was performed by incrementally updating the model parameters using a single image-encoding pair from the data owner's original dataset $\{X_A, Z_A\}$ and then gradually including more

image-encoding pairs in the training process until SSIM saturates. The model parameters are updated by minimizing the mean squared error between original images and reconstructed images given the encoded samples. After each training step, the image similarity of the data owner's original images and the reconstructed images are measured using SSIM.

5.1.5 CAE Minimal Data Subset Attack

In the CAE experiments, this study considers the scenario where an attacker trains a randomly initialized (RI) decoder network using a subset of the data owner's original image-encoding pair. Also, this study considers the scenario where the attacker first pre-trains a decoder model using his constructed dataset X_B and then fine-tunes the decoder (FT) using a subset of the data owner's original image-encoding pair.

5.1.6 ViT Minimal Data Subset Attack

In the ViT experiments, this study considers the scenario where an attacker trains a randomly initialized MAE network using a subset of the original image-encoding pairs. Also, this study considers the scenario where an attacker first pre-trains an MAE network using the constructed dataset and then fine-tunes the network using the subset of original image-encoding pairs. Similar to previous experiments the network parameters are learned using the mean squared error loss. Afterwards, the CAE and MAE decoders are used to reconstruct the data owner's encoded dataset.

5.1.7 Reconstruction Cycle GAN Attack

Next, this study considers the scenario where an attacker attempts to reconstruct the data owner's original dataset by estimating the encoding function using a Cycle Gan based approach. Cycle GANs are normally used to learn a mapping function from one data domain to another [113]. In

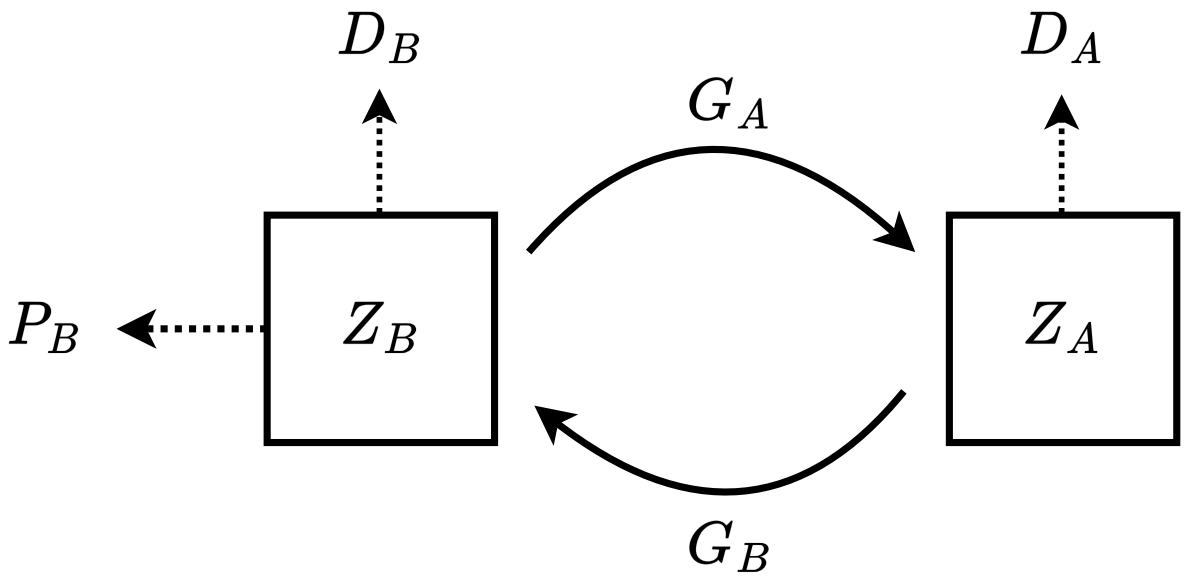


Figure 5.1: Cycle GAN reconstruction attack diagram. Where Z_B is the adversaries encoded dataset and Z_A is the data owner encoded set. Generator G_A translates the adversaries encoded set into the data owners encoded set domain. Generator G_B translates the data owner's encoded set into the adversaries encoded set domain. Discriminator D_B distinguishes between true adversary encoded images and fake adversary encoded images. Discriminator D_A distinguishes between true data owner encoded images and fake data owner encoded images. Decoder P_B reconstructs the adversaries encoded images.

this work, the goal is to learn a mapping function between the data owner's encoded set and the attacker's encoded set. This work assumes an attacker only has access to the encoded dataset and corresponding labels.

In this work, the reconstruction Cycle GAN attack network consists of two generators (G_A, G_B), two discriminators (D_A, D_B) and a pre-trained decoder (P_B). The attacker's encoded dataset is generated using pre-trained encoding function $Z_B = E_B(X_B)$. Generator G_A is used to translate encoded images from the attacker's domain to the data owner's domain and generator G_B is used to translate encoded images from the data owner's domain to the attacker's domain. Discriminator D_A is used to distinguish between the data owner's real and fake encoded samples and discriminator D_B is used to distinguish between the attacker's real and fake encoded samples. The pre-trained decoder P_B is used to reconstruct the attacker's dataset given the encoded set as depicted in Figure 5.1.

5.1.8 Reconstruction Cycle GAN Adversarial Loss

In this work, the adversarial loss term is computed using generator G_A, generator G_B, discriminator D_A and discriminator D_B. Discriminator D_A is a binary classifier used to distinguish between the data owner's real encoded set Z_A and the generated encoded Z'_A. First, generator G_A is used as a mapping function from the attacker's encoded image domain to the data owner's encoded image domain $Z'_A = G_A(Z_B)$. Generator G_A wishes to minimize the probability of Z'_A being classified as a generated data sample by discriminator D_A while D_A aims to maximize the probability of the real encoded datatset Z_A being classified as real encodings and generated encodings Z'_A being classified as fake encodings. The attacker learns a generator G_A that translates encoded samples Z_B into the data owner's domain.

Discriminator D_B is a binary classifier used to distinguish between real and generated attacker encodings. This study simulates the process of the data owner generating the attacker's encoded images to obtain Z'_B using generator G_B given the data owner's encoded set as input to generator G_B, i.e. $Z'_B = G_B(Z_A)$. Generator G_B wishes to minimize the probability of Z'_B being classified as a generated encodings by discriminator D_B while D_B aims to maximize the probability of true encodings Z_B being classified as a true encodings and generated encodings Z'_B being classified as fake encodings. The simulated data owner learns a generator that translates her encoded set into the attacker's domain.

The full adversarial loss consists of a loss term from generators (G_A, G_B) and discriminators (D_A, D_B). The following equations describe the adversarial loss term.

$$\begin{aligned} L_{GAN}(G_A, D_A, Z_B, Z_A) = \mathbb{E}_{z_a \sim p_a(z_a)}[\log D_A(z_a)] \\ + \mathbb{E}_{z_b \sim p_b(z_b)}[\log(1 - D_A(G_A(z_b)))] \end{aligned} \tag{5.1}$$

where G_A tries to generate encodings $G_A(z_b)$ that are similar to the data owner's encoded

images z_a, while D_A distinguishes between real data owner encodings z_a and generated data owner encodings $G_A(z_b)$. G_A minimizes the objective while D_A maximizes the objective, $\min_{G_A} \max_{D_A} L_{GAN}(G_A, D_A, Z_B, Z_A)$.

$$\begin{aligned} L_{GAN}(G_B, D_B, Z_A, Z_B) = \mathbb{E}_{z_b \sim p_b(z_b)}[\log D_B(z_b)] \\ + \mathbb{E}_{z_a \sim p_a(z_a)}[\log(1 - D_B(G_B(z_a)))] \end{aligned} \tag{5.2}$$

where G_B tries to generate encodings $G_B(z_a)$ that are similar to the attacker's encodings z_b, while D_B distinguishes between the attacker's real encodings and generated $G_B(z_a)$ encodings. G_B minimizes the objective while D_B maximizes the objective, $\min_{G_B} \max_{D_B} L_{GAN}(G_B, D_B, Z_A, Z_B)$.

5.1.9 Reconstruction Cycle GAN Cycle Consistency Loss

The cycle consistency loss term is computed using generator G_A and generator G_B. First, the attacker translates his encoded set Z_B into the data owner's domain using generator G_A. Then the encoding is translated back into the attacker's data domain using generator G_B. Second, the simulated data owner translates her encoded set Z_A into the attacker's domain using generator G_B. Then the encoded set is translated back into data owner's domain using generator G_A. The mean absolute error between the original encoded set and the cycled encoded set are computed.

The computed cycle consistency loss values for the cycled encoded data are summed together below.

$$\begin{aligned} L_{cyc}(G_A, G_B) = \mathbb{E}_{z_b \sim p_b(z_b)}[||G_B(G_A(z_b)) - z_b||_1] \\ + \mathbb{E}_{z_a \sim p_a(z_a)}[||G_A(G_B(z_a)) - z_a||_1] \end{aligned} \tag{5.3}$$

$G_B(G_A(z_b))$ is the attacker's cycled encoding and $G_A(G_B(z_a))$ is the data owner's cycled encoding. The error between the cycled encoding and real encoding is minimized and combined to compute the total cycle consistency loss.

Next, the attacker reconstructs his original image dataset from the encoded set using his pre-trained decoder $X'_B = P_B(Z_B)$. Then a distorted version of the attacker's original image dataset is reconstructed from his cycled encoded set $\hat{X}'_B = P_B(G_B(G_A(Z_B)))$. The mean absolute error between the reconstructed image dataset X'_B and the reconstructed distorted version of the image dataset $\hat{X}'_B$ is computed.

The reconstruction loss L_r minimizes the error between the reconstructed dataset given the attacker's encoded set and the reconstructed dataset given the attacker's cycled encoded set as described below.

$$L_r(P_B, G_B, G_A) = \mathbb{E}_{z_b \sim p_b(z_b)}[||P_B(G_B(G_A(z_b))) - P_B(z_b)||_1] \tag{5.4}$$

where P_B is the attacker's pre-trained decoder. Decoder P_B is fine-tuned by minimizing $||P_B(G_B(G_A((Z_B))) - P_B(Z_B)||_1$ to ensure that the reconstructed cycled encodings remain similar to the reconstructed encodings.

5.1.10 Reconstruction Cycle GAN Attack Full Objective

All of the previously discussed loss terms are summed together for the full objective. The full objective for the attack consists of an adversarial loss term, cycle consistency loss term and reconstruction loss term.

The full objective is:

$$\begin{aligned} L(G_A, G_B, D_B, D_A, P_B) = L_{GAN}(G_A, D_A, Z_b, Z_a) \\ + L_{GAN}(G_B, D_B, Z_a, Z_b) \\ + \lambda_1 L_{cyc}(G_A, G_B) \\ + \lambda_2 L_r(P_B, G_B, G_A) \end{aligned} \tag{5.5}$$

where λ controls the importance of each objective. In the experiments, $\lambda_1 = 10$ and $\lambda_2 = 10$. This study solves the following optimization problem:

$$\begin{aligned} G_A^*, G_B^*, P_B^* = \\ \arg \min_{G_A, G_B, P_B} \max_{D_A, D_B} L(G_A, G_B, D_B, D_A, P_B) \end{aligned} \tag{5.6}$$

5.2 Image Anonymization Evaluation

* The material presented in this section previously appeared in the proceedings of the 37th Annual IFIP WG 11.3 Conference on Data and Applications Security and Privacy (DBSec'23) in the article, "An Autoencoder-Based Image Anonymization Scheme for Privacy Enhanced Deep Learning", co-authored with Ram Krishnan, Ph.D.

5.2.1 Evaluating the Privacy/Utility Trade-off

The anonymization network is trained using the proposed method and afterward this study examines the trade-off between privacy and utility, i.e. we measure the change in identity and attribute classification accuracy. First, the original images are transformed using the proposed anonymization method. Second, the identity and attribute classification accuracy of original images and the transformed images are compared. To quantify the trade-off between privacy and utility this study

measures the difference in identity and attribute classification accuracy for the anonymized dataset compared to original dataset.

5.2.2 Image Anonymization Evaluation with Classifier Transfer Attack

This study evaluates the robustness of the autoencoder-based image anonymization approach against attacks that aim to learn an identity feature classifier and transfer it onto the data owners encoded set for classification. This study conducts experiments on CelebA and Cifar-100 datasets using gender and coarse labels, respectively. This study assumes that the attacker is able to construct a dataset that follows a similar probability distribution as the data owner's original dataset and corresponding labels. First, the attacker trains his own identity classifier using the constructed dataset to achieve high classification accuracy. Then he attempts to classify the data owners encoded set using his pre-trained identity classifier. An overview of the classifier transfer attack is depicted in Figure 5.2.

The performance of the attacker's pre-trained identity classifier is evaluated using the data owner's encoded set. The goal of the attack is to classify identity features given the data owner's encoded dataset.

5.2.3 Image Anonymization Evaluation with Encoding Transfer Attack

This study also considers the scenario where the attacker aims to learn a representation of the data owner's encoded set to classify identity features. Again, this study assumes the attacker constructs a dataset that follows a similar distribution as the data owner's original dataset and corresponding labels. First, the attacker trains a multi-output classification model for identity and attribute features similar to the proposed method. Then a randomly initialized autoencoder is trained to generate encoded samples such that identity and attribute information are both preserved. This is accomplished by freezing the weights of the pre-trained identity and attribute classifier

and updating the autoencoder parameters based on the gradients of the classification loss. The attacker's modified anonymization network is trained to maintain high classification accuracy for both the identity and attribute classifiers. The encoding transfer attack is depicted in Figure 5.3.

Finally, the effectiveness of the proposed method evaluated against encoding transfer attacks by assessing the performance of the data owner's identity classifier given the attacker's generated encoded set. The goal of the attack is to generate encoded images that include exploitable identity features. The data owner's identity classifier is used to verify if identity features are present in the attackers encoded set.

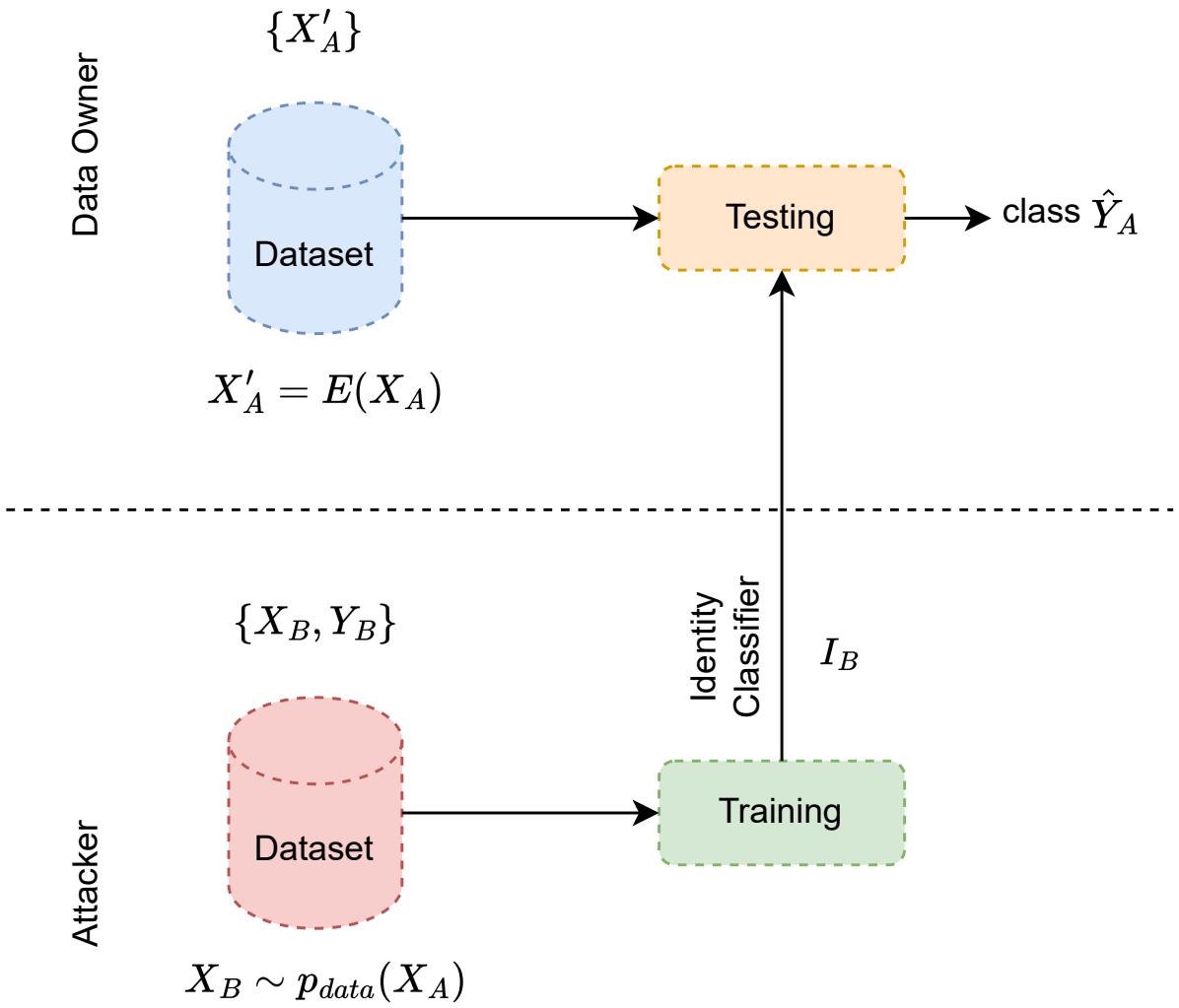


Figure 5.2: Classifier transfer attack diagram. Where X'_A is the data owner's encoded dataset and X_B, Y_B are the attackers raw image dataset and identity labels which follows the probability distribution of the data owner's original dataset. I_B is the attacker's identity classifier. The attacker trains I_B with X_B, Y_B and uses the classifier to predict the identity label of the data owner's encoded dataset.

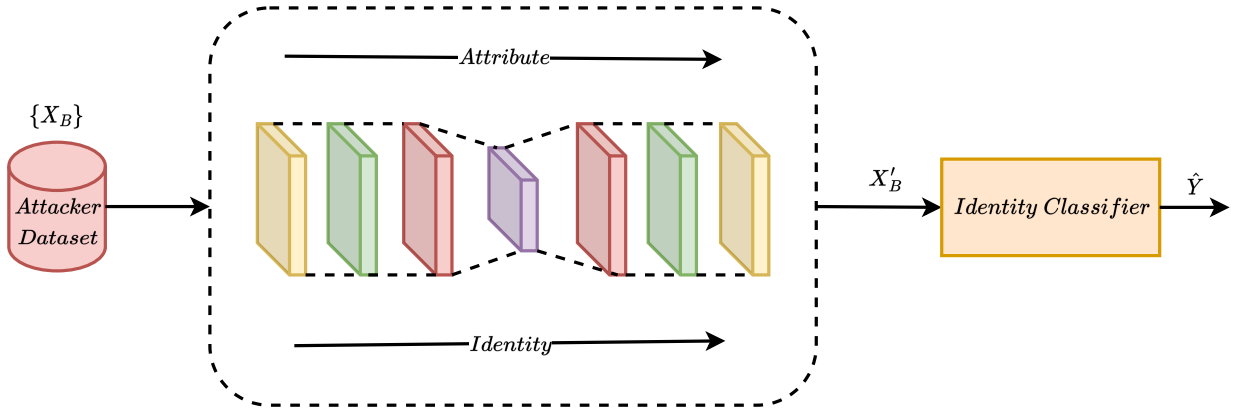


Figure 5.3: Encoding transfer attack diagram. Where X_B is the attacker's dataset which follows the probability distribution of the data owner's original dataset. X'_B is the attacker's encoded dataset which consists of attribute features and identity features. The data owners identity classifier is used to predict the identity label of the attacker's encoded dataset to verify if X'_B captures the data owner's identity features.

5.3 Adversarial Robustness Evaluation

* The material presented in this section previously appeared in the proceedings of the International Conference on Intelligent Biology and Medicine (ICIBM 2021) in the article, "On the role of deep learning model complexity in adversarial robustness for medical images", co-authored with Tapsya Nayak, Ph.D., et al.

5.3.1 Adversarial Robustness & Model Complexity

DNN models of sufficient complexity were developed to achieve low train and test error as shown in Figure 5.4 Adversarial examples were crafted for each of the trained DNNs and model performance was assessed by obtaining the average accuracy for each experiment. The performance of each network was evaluated as model complexity was reduced and the magnitude of perturbation increased as depicted in Figure 5.5.

5.3.2 Adversarial Attack Evaluation

The performance curves include results for both FGSM and PGD attacks but more emphasis will be placed on results for PGD attacks since it is stronger variant. This study considers a model to be more robust to an adversarial attack if the amount of perturbation that is required to degrade the model's performance is greater than other networks. Attacks were launched on each model with a gradually increasing perturbation to assess the degradation performance incurred for the system under attack. This study considers networks to be less robust if less perturbation is required to degrade the model's performance. The networks that required a higher degree of perturbation to confidently change the model's output prediction were considered more robust.

The medical datasets contain a small amount of classes so this study investigates whether the number of classes contributes to the findings on medical DNNs. In doing so, this study repli-

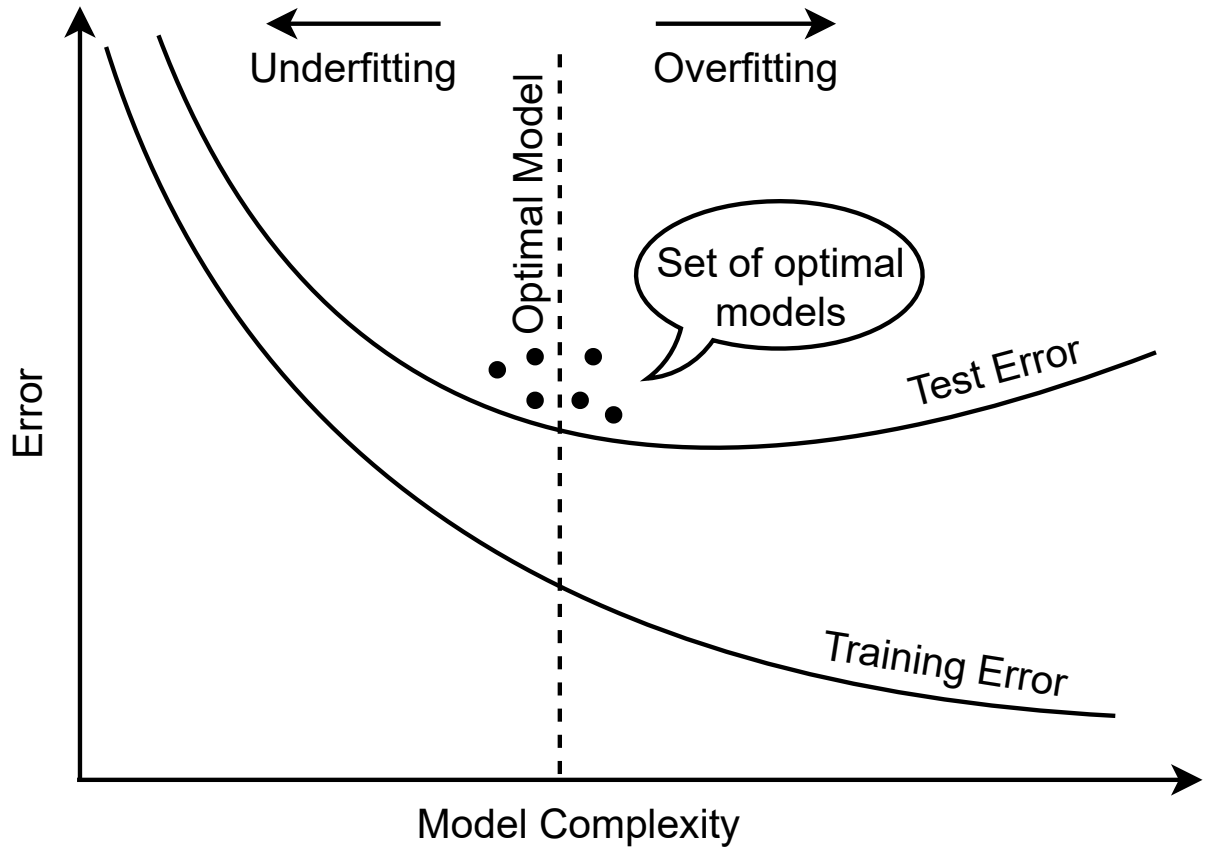


Figure 5.4: Set of optimal models with sufficient complexity for generalization. The set of optimal models must achieve low train and test error.

cates experiments using cifar10 and mnist datasets while also reducing the number of classes down equivalently to the medical datasets which have 2 classes for x-ray and melanoma and 4 classes for OCT dataset. It is typical for medical datasets to contain far less classes than natural image dataset and far less data samples.

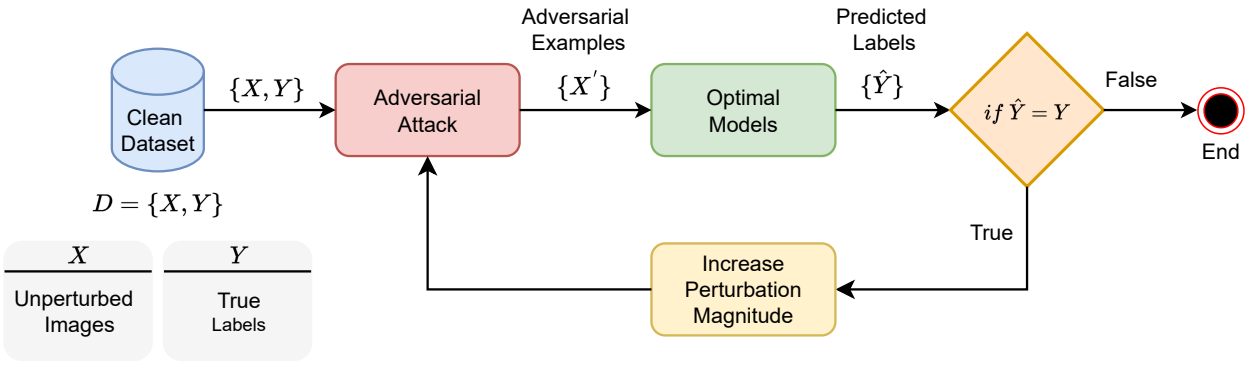


Figure 5.5: Adversarial robustness for a given set of optimal models of sufficient complexity. Adversarial examples are generated and all optimal models are attacked with increasing perturbation magnitude until model performance degrades.

CHAPTER 6: EXPERIMENTAL RESULTS

6.1 CAE and ViT Image Encoding Results

* The material presented in this section is currently being reviewed to appear in the proceedings of the 10th European Conference On Service-Oriented And Cloud Computing (ESOCC 2023) in the article, "Evaluating Robustness of CAE and ViT Image Encoding for Privacy Enhanced Image Classification", co-authored with Ram Krishnan, Ph.D. and Yufei Huang, Ph.D.

The CAE and ViT encoding schemes enhance the privacy of sensitive image data while preserving classification accuracy using Fashion Mnist, Cifar-10 and Chest X-ray datasets.

6.1.1 Encoded Image Classification Model Performance

First, the DNN latent space classifier and ViT embedding classifier are trained using the data owner's encoded set $\{z_{a_i}\}_{i=1}^{N}$ and corresponding class labels $\{y_{a_i}\}_{i=1}^{N}$. Then, this study evaluates the performance of each network using a test dataset of encoded images that were held out of the training process. The classification accuracy of the DNN latent space classifier shown in row 1 of Table 6.4 is 91.43%, 89.94% and 92.67% for Fasion Mnist, Cifar-10 and Chest X-ray datasets, respectively. The classification accuracy of the ViT embedding classifier shown in row 2 of Table 6.4 is 87.32%, 86.01% and 91.98% for Fasion Mnist, Cifar-10 and Chest X-ray datasets, respectively. High classification accuracy is obtained on encoded images using the DNN latent space classifier and ViT embedding classifier. This result demonstrates that model utility remains high using encoded datasets.

Table 6.1: Privacy enhanced image classification accuracy. Model utility is preserved for DNN and ViT classifiers trained using encoded datasets.

	Classification Acc.%		
Privacy Enhanced Classifier	Fashion Mnist	Cifar-10	Chest X-ray
Latent Space Classifier (DNN)	91.43	89.94	92.67
Embedding Classifier (ViT)	87.32	86.01	91.98

6.1.2 Public/Query Encoder Attack Results

The performance of the public and query encoder attack are evaluated using structural similarity index measure (SSIM). The SSIM values that are closer to 1 indicate that the reconstructed images are similar to the original images and values closer to 0 indicate that reconstructed images are poor quality compared to original images. The public encoder attack SSIM scores for CAE and ViT image reconstruction are shown in row 1 of Table 6.5 and Table 6.3, respectively. The CAE public encoder attack SSIM scores for Fashion Mnist, Cifar-10 and Chest X-ray datasets are 0.9385, 0.9012 and 0.9157, respectively. The ViT public encoder attack SSIM scores for Fashion Mnist, Cifar-10 and Chest X-ray datasets are 0.1968, 0.3598 and 0.2413, respectively. The query encoder attack SSIM scores for CAE and ViT image reconstruction are shown in row 2 of Table 6.5 and Table 6.3, respectively. The CAE query encoder attack SSIM scores for Fashion Mnist, Cifar-10 and Chest X-ray datasets are 0.9258, 0.9127 and 0.9036, respectively. The ViT query encoder attack SSIM scores for Fashion Mnist, Cifar-10 and Chest X-ray datasets are 0.1966, 0.3587 and 0.2474, respectively. The high quality CAE image reconstruction is due to the attacker's accessibility to the data owner's original encoding function. The low ViT SSIM score indicates poor image reconstruction quality. The public and query encoder attacks are a baseline attack methods with the strong assumption that an attacker has access to the data owner's original encoding function.

Table 6.2: CAE image reconstruction attack SSIM results. SSIM scores near 1 indicate high quality image reconstruction whereas scores closer to 0 indicate poor quality imge reconstruction.

		CAE SSIM Score		
Attack Method	Attacker's Knowledge	Fashion Mnist	Cifar-10	Chest X-ray
Public Encoder	E_A, Z_A, Y_A	0.9385	0.9012	0.9157
Query Encoder	E_A, Z_A, Y_A	0.9258	0.9127	0.9036
Min. Data Subset (FT)	X_A, Z_A, Y_A	0.8871	0.8726	0.8533
Min. Data Subset (RI)	X_A, Z_A, Y_A	0.8580	0.8692	0.8369
Cycle GAN Recon.	Z_A, Y_A	0.1984	0.1727	0.1615

6.1.3 Minimal Data Subset Attack Results

This study evaluates the performance of minimal data subset attacks using SSIM. In the results, this study reports similarity scores as SSIM begins to saturate. The minimal data subset attack SSIM scores for fine-tuned CAE and ViT image reconstruction are shown in row 3 of Table 6.5 and Table 6.3, respectively. The fine-tuned CAE SSIM scores for Fashion Mnist, Cifar-10 and Chest X-ray datasets are 0.8871, 0.8726 and 0.8533, respectively. The fine-tuned ViT SSIM scores for Fashion Mnist, Cifar-10 and Chest X-ray datasets are 0.2201, 0.3745 and 0.2618, respectively. The minimal data subset attack SSIM scores for randomly initialized CAE and ViT image reconstruction are shown in row 4 of Table 6.5 and table 6.3, respectively. The randomly initialized CAE SSIM scores for Fashion Mnist, Cifar-10 and Chest X-ray datasets are 0.8580, 0.8692 and 0.8369, respectively. The randomly initialized ViT SSIM scores for Fashion Mnist, Cifar-10 and Chest X-ray datasets are 0.2295, 0.3778 and 0.2601, respectively. The CAE SSIM scores are indicative of high quality image reconstruction which are the result of the attacker's ability to access a subset of the original image-encoding pairs. This result informs that only a fraction of the data owner's original image-encoding pairs are required to reconstruct X_A given Z_A. The ViT SSIM scores are indicative of low quality image reconstruction.

Table 6.3: ViT image reconstruction attack SSIM results. SSIM scores near 1 indicate high quality image reconstruction whereas scores closer to 0 indicate poor quality imge reconstruction.

		ViT SSIM Score		
Attack Method	Attacker's Knowledge	Fashion Mnist	Cifar-10	Chest X-ray
Public Encoder	E_A, Z_A, Y_A	0.1968	0.3598	0.2413
Query Encoder	E_A, Z_A, Y_A	0.1966	0.3587	0.2474
Min. Data Subset (FT)	X_A, Z_A, Y_A	0.2201	0.3745	0.2618
Min. Data Subset (RI)	X_A, Z_A, Y_A	0.2295	0.3778	0.2601
Cycle GAN Recon.	Z_A, Y_A	0.1197	0.1352	0.1092

6.1.4 Reconstruction Cycle GAN Attack Results

This work demonstrates an attacker's ability to reconstruct the data owner's original image dataset using the reconstruction Cycle GAN attack method. As previously mentioned, the attacker's pre-trained decoder is fine-tuned by minimizing the error between the reconstructed dataset given his encoded set i.e., $P_B(Z_B)$ and the reconstructed dataset given the cycled version of his encoded set i.e., $P_B(G_B(G_A(Z_B)))$. Decoder P_B^* was optimized to reconstruct the attacker's image data using cycled encoded samples. The reconstructed images are expected to consist of inherent features from the data owner's domain as Cycle GAN learns a mapping from one domain to another. Thus, the data owner's encoded set is translated to the attacker's domain $G_B(Z_A)$ to reconstruct the translated encoding using decoder P_B^*, i.e., $P_B^*(G_B(Z_A))$. The SSIM score between the reconstructed images and the original images are shown in row 5 of Table 6.5 and Table 6.3 for CAE and ViT, respectively. The SSIM scores report using X_A and $P_B^*(G_B(Z_A))$ for Fashion Mnist, Cifar-10 and Chest X-ray datasets. In the reconstruction Cycle GAN experiments, this study demonstrates that image reconstruction is of poor quality given that an attacker only has access to the encoded set and corresponding labels. Thus, if an attacker's knowledge is limited to $\{Z_A, Y_A\}$ then reconstructed images are most dissimilar to the data owner's original private image dataset when compared to all other attack methods.

Table 6.4: Image Classification Accuracy of identity and attribute classifier for CelebA and Cifar-100 datasets

	Identity Acc (%)		Attribute Acc (%)	
Encryption	CelebA	Cifar-100	CelebA	Cifar-100
Plain Images	95.85	82.96	87.02	88.13
Proposed Scheme	50.33	20.71	85.96	83.45

6.2 Image Anonymization Results

In the results, this study demonstrates that the image anonymization method increases data privacy while maintaining model utility using CelebA [57] and Cifar-100 [**?**] datasets. The identity classification accuracy of anonymized images significantly decreased compared to original images. Additionally, the attribute classification accuracy of anonymized images is similar to original images. To quantify the trade-off between privacy and utility the reduction in identity and attribute classification accuracy for the anonymized dataset compared to original dataset is measured. In the experiments, this study demonstrates that the proposed image anonymization method enables us to maintain high image attribute classification accuracy of 85.96% & 83.45% for CelebA and Cifar-100 datasets, respectively, which is similar to original images. It also enables us to reduce image identity classification accuracy from 95.85% & 82.96% to 50.33% & 20.71% for CelebA and Cifar-100 datasets, respectively, as shown in Table 6.4. The identity classification accuracy of anonymized images significantly decreased compared to original images. Additionally, the attribute classification accuracy of anonymized images is similar to original images.

6.2.1 Classifier and Encoder Transfer Attack Results

The experimental results of the proposed image encoding method against classifier transfer attacks is assessed using the attackers pre-trained identity classifier. The performance of the attacker's pre-trained identity classifier is evaluated using the data owner's encoded set. The goal of the attack is to classify identity features given the data owner's encoded dataset. The experimental results

Table 6.5: Classifier and encoding transfer attack performance on CelebA and Cifar-100 datasets

	Identity Acc (%)	
Attack Scheme	CelebA	Cifar-100
Classifier Transfer	23.49	17.01
Encoding Transfer	25.59	20.58

demonstrate that the proposed method is resistant against classifier transfer attacks as shown in row 1 of Table 6.5 the classification accuracy is 23.49% and 17.01% for CelabA and Cifar-100, respectively.

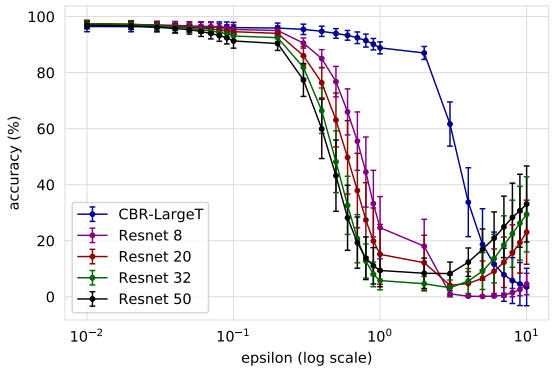
Finally, the effectiveness of the proposed method against encoding transfer attacks is evaluated by assessing the performance of the data owner's identity classifier given the attacker's generated encoded set. The goal of the attack is to generate encoded images that include exploitable identity features. The data owner's identity classifier is used to verify if identity features are present in the attackers encoded set. The experimental results show that the proposed method is resistant to encoding transfer attacks as shown in row 2 of Table 6.5, the classification accuracy is poor 25.59% and 20.58% for CelebA and Cifar-100, respectively.

6.3 Adversarial Robustness Results

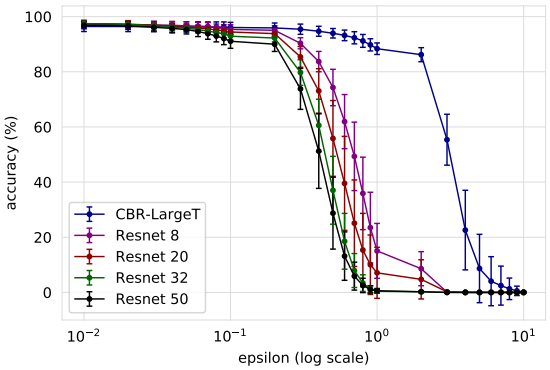
This research focuses on the magnitude of perturbation that is required to cause the networks to incorrectly classify a given input sample. The goal of an attacker is to utilize an imperceptible perturbation to successfully attack the model, as a result, this study bases its results on the least amount of perturbation that causes the highest drop in accuracy. In Figure 6.1 there is a significant drop in accuracy for ϵ = 1, 0.1 and 1 for chest x-ray, dermoscopy, and OCT, respectively. Models with lower complexity demonstrate greater robustness. Tables 6.6 - 6.8 provide the average accuracy values for each ϵ that corresponds to the least perturbation with the largest margin between the least and most robust networks.

These experiments reveal an inverse relationship between model complexity and adversarial robustness for medical and cifar10 datasets. In particular, this study verifies that reduced model complexity is crucial for adversarial robustness on medical and natural image DNNs. Medical and natural image DNNs seem to benefit more from less complex networks, in terms of adversarial robustness. In other words, as the level of data complexity increases, the level of model complexity should decrease for adversarial robustness to be accomplished.

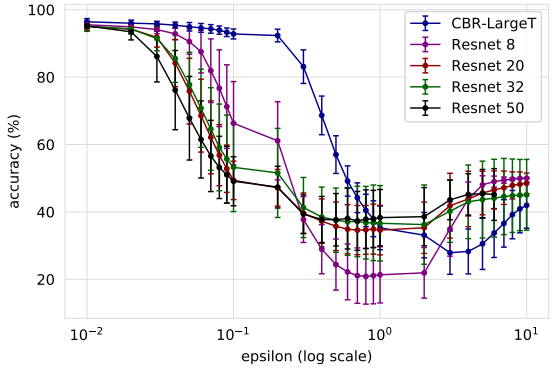
As the perturbation increased the average accuracy decreased for all models when epsilon is relatively large. Most importantly, for models trained on medical images and cifar10 we observed that as each model was attacked with a range of increasing perturbation sizes and decreasing architecture complexity, the adversarial robustness increases for many perturbations.



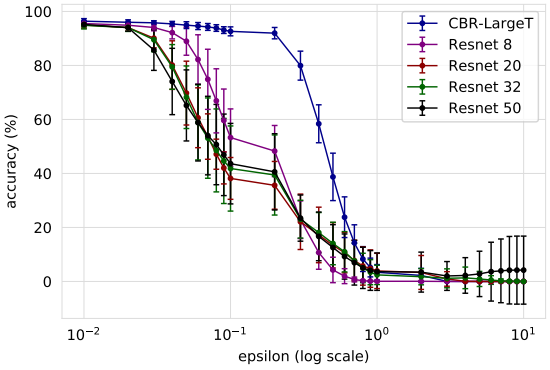
(a) Chest X-Ray, FGSM



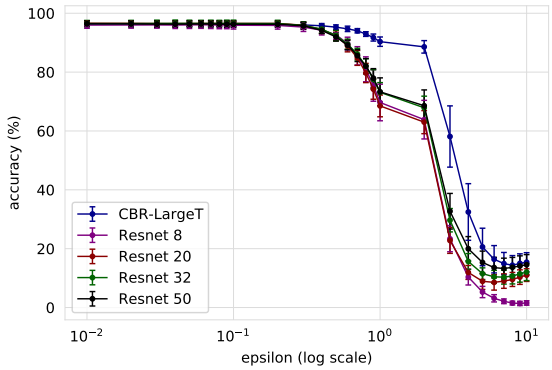
(b) Chest X-Ray, PGD



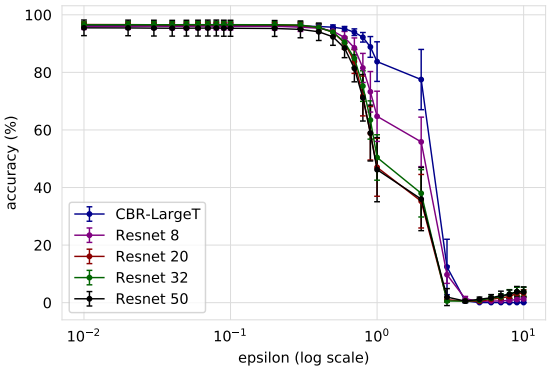
(c) Dermoscopy, FGSM



(d) Dermoscopy, PGD



(e) Optical Coherence Tomography, FGSM



(f) Optical Coherence Tomography, PGD

Figure 6.1: The average accuracy and standard deviation of adversarial attacks on medical images. For all medical datasets the models of reduced complexity exhibit greater adversarial robustness, this is especially true for the PGD attacks. All networks exhibit similar performance on unperturbed data.

Attack	CBR-LargeT	Resnet-8	Resnet-20	Resnet-32	Resnet-50
No Attack	96.43	97.43	97.46	97.41	96.90
FGSM	88.83	24.63	15.13	5.77	9.37
PGD	88.37	15.07	7.10	0.43	0.60

Table 6.6: Chest X-Ray Average Accuracy, $\epsilon = 1$.

Attack	CBR-LargeT	Resnet-8	Resnet-20	Resnet-32	Resnet-50
No Attack	96.40	95.53	95.03	95.07	95.20
FGSM	92.80	66.30	49.37	53.23	49.10
PGD	92.63	53.30	38.17	41.83	43.63

Table 6.7: Dermoscopy Average Accuracy, $\epsilon = 0.2$.

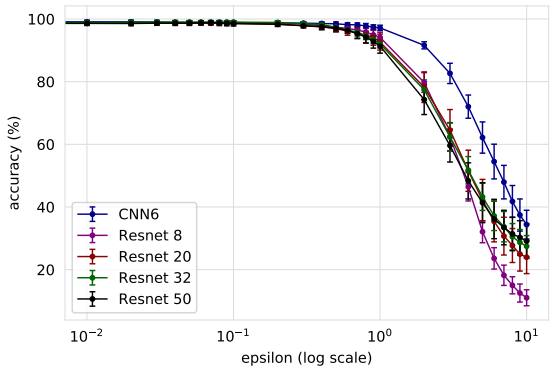
Attack	CBR-LargeT	Resnet-8	Resnet-20	Resnet-32	Resnet-50
No Attack	96.30	95.53	95.03	95.07	95.20
FGSM	88.58	63.85	63.00	67.88	68.65
PGD	77.75	55.85	35.23	37.98	35.98

Table 6.8: OCT Average Accuracy, $\epsilon = 2$.

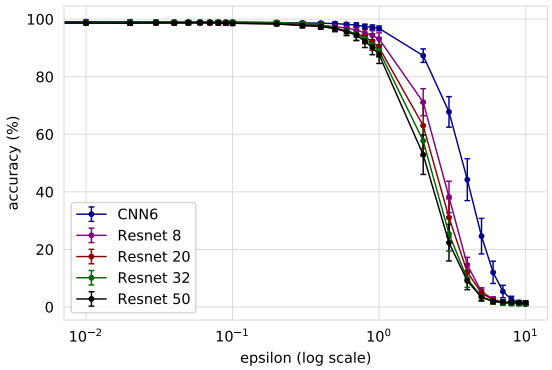
6.3.1 Cifar10 Model Performance

For the cifar10 experiments there is a similar trend for all versions of the dataset including 10 classes, 4 classes and 2 classes. As previously mentioned study reduces the number of classes to represent the amount of labels seen in the medical datasets. The original 10 class version of cifar10 exhibits a slight reduction in performance for clean data and negligible perturbations. In this case, there is a trade off between accuracy and robustness. After the magnitude of perturbation $\epsilon = 1$ the standard CNN model is consistently more robust than any other models. In all other experiments the models of reduced complexity maintained a greater degree of adversarial robustness. In particular, the standard CNN model of reduced complexity was consistently more robust than all other models. It is typical for practitioners to utilize very large architectures in the development of deep learning algorithms in many domains but the following results demonstrate that this practice

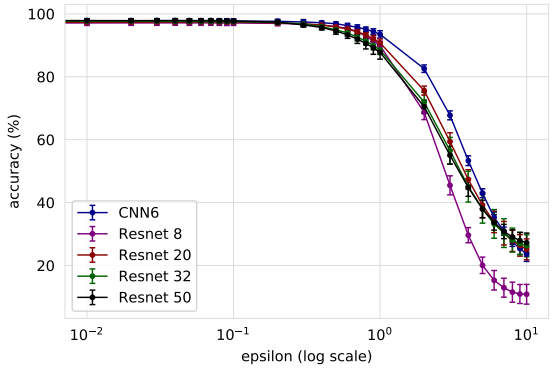
may not always be the most secure route when designing robust networks. In Figure 6.2 there is a significant drop in accuracy for $\epsilon = 2$ for cifar2, cifar4, and cifar10. Models with lower complexity demonstrate greater robustness for the cifar10 dataset. Tables 6.9 - 6.11 provide the average accuracy values for each ϵ that corresponds to the least perturbation with the largest margin between the least and most robust networks.



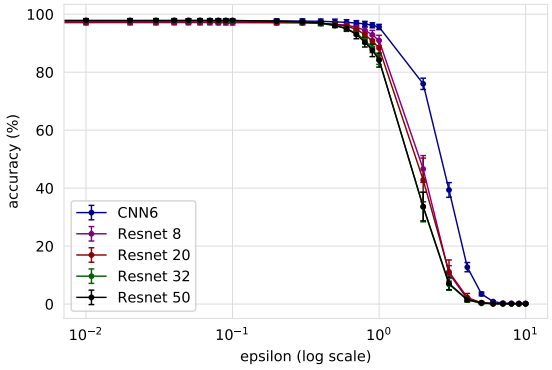
(a) Cifar2, FGSM



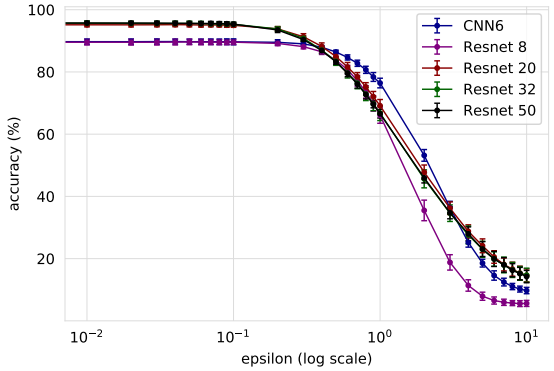
(b) Cifar2, PGD



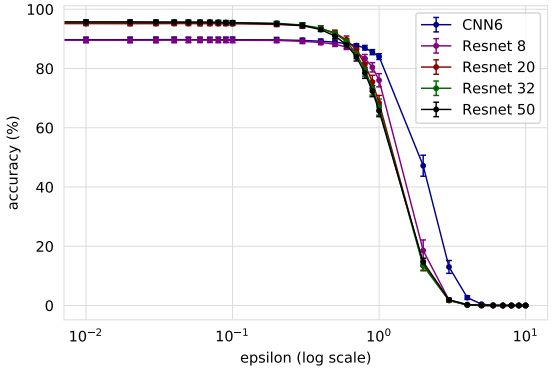
(c) Cifar4, FGSM



(d) Cifar4, PGD



(e) Cifar10, FGSM



(f) Cifar10, PGD

Figure 6.2: The average accuracy and standard deviation of adversarial attacks on the Cifar10 datasets. For all versions of the Cifar10 datasets the models of reduced complexity exhibit greater adversarial robustness, this is especially true for the PGD attacks. Note that there is a small tradeoff between between accuracy and robustness for cifar10 as the models of lowest complexity generate slightly lower performance on unperturbed data but offer greater robustness prior to $\epsilon = 1$ for PGD attack.

Attack	CNN6	Resnet-8	Resnet-20	Resnet-32	Resnet-50
No Attack	99.17	98.77	98.73	98.87	98.53
FGSM	91.60	79.73	78.23	77.4	74.27
PGD	87.30	71.13	62.99	57.73	52.90

Table 6.9: Cifar2 Average Accuracy, $\epsilon = 2$

Attack	CNN6	Resnet-8	Resnet-20	Resnet-32	Resnet-50
No Attack	97.73	97.10	97.32	97.67	97.87
FGSM	82.57	68.70	75.60	71.92	70.47
PGD	75.98	46.60	42.81	33.48	33.63

Table 6.10: Cifar4 Average Accuracy, $\epsilon = 2$

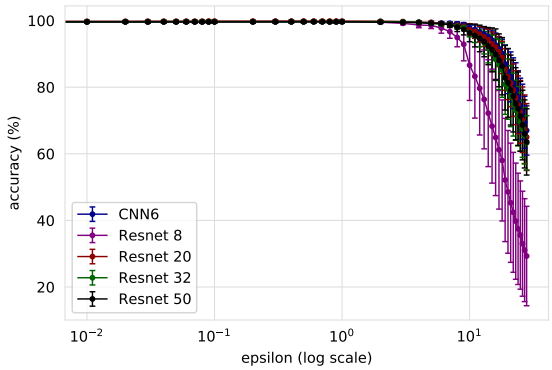
Attack	CNN6	Resnet-8	Resnet-20	Resnet-32	Resnet-50
No Attack	89.73	89.56	95.15	95.63	95.75
FGSM	76.43	66.10	69.15	66.73	66.56
PGD	84.03	76.04	68.41	66.96	65.69

Table 6.11: Cifar10 Average Accuracy, $\epsilon = 1$

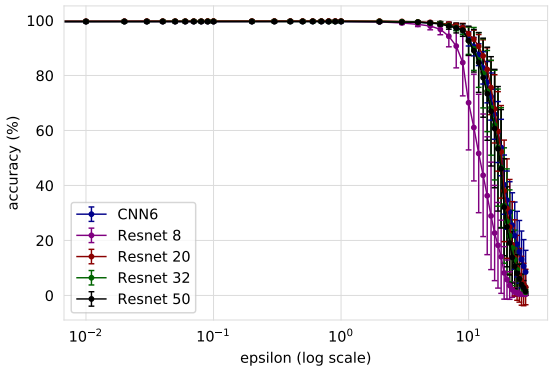
6.3.2 Mnist Model Performance

The mnist dataset results were fairly consistent for most networks providing similar performance and robustness across all models and datasets with the exception of mnist4 where the networks of reduced complexity offers greater robustness and similar performance on unperturbed data. One thing to note is that the mnist dataset requires much more perturbation to degrade model performance. This is surprising since previous works have suggested that one possible explanation for medical datasets being easier to attack (requiring less perturbation to degrade model performance) is due to the simplicity of the data, if this was the case then the mnist dataset should require significantly less perturbation to degrade model performance but the current results demonstrate the opposite. In Figure 6.3 there is not a significant drop in accuracy for mnist2 and mnist10 but there is a drop in performance for mnist4 beginning prior to $\epsilon = 1$. Models with lower complexity do not

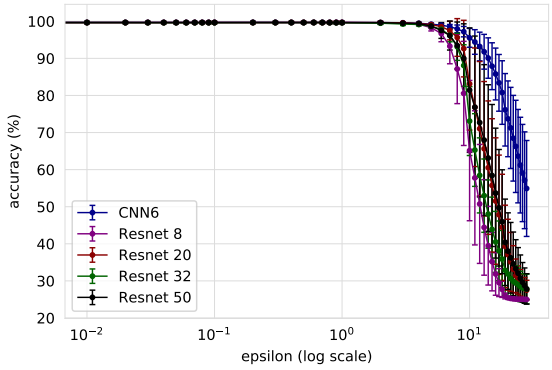
demonstrate greater robustness for mnist2 and mnist10 but they do offer comparable performance to more complex networks. Tables 6.9 - 6.11 provide the average accuracy values for each ϵ that corresponds to the least perturbation with the largest margin between the least and most robust networks.



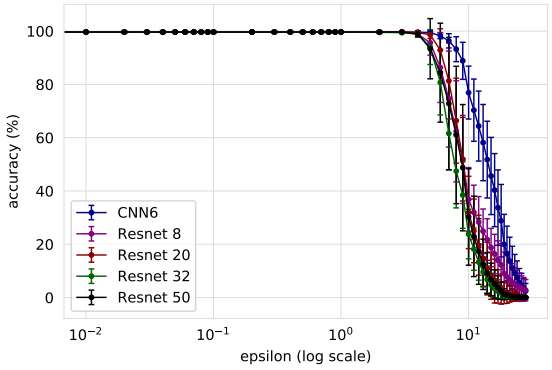
(a) Mnist2, FGSM



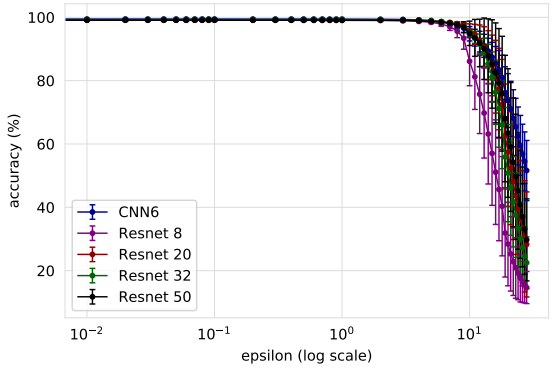
(b) Mnist2, PGD



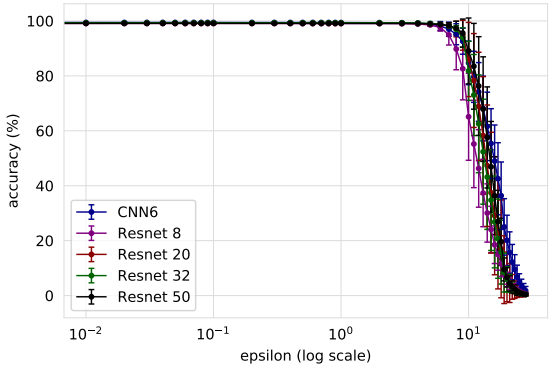
(c) Mnist4, FGSM



(d) Mnist4, PGD



(e) Mnist10, FGSM



(f) Mnist10, PGD

Figure 6.3: The average accuracy and standard deviation of adversarial attacks on Mnist datasets. The adversarial robustness of mnist2 and mnist10 are mainly constant across all models. The mnist4 dataset demonstrates greater robustness for the model of reduced complexity while achieving comparable performance on unperturbed data.

Attack	CNN6	Resnet-8	Resnet-20	Resnet-32	Resnet-50
No Attack	99.67	99.60	99.78	99.63	99.60
FGSM	99.03	94.93	98.63	98.30	98.03
PGD	97.63	90.70	98.13	97.40	97.27

Table 6.12: Mnist2 Average Accuracy, ϵ = 10

Attack	CNN6	Resnet-8	Resnet-20	Resnet-32	Resnet-50
No Attack	99.70	99.63	99.67	99.65	99.65
FGSM	97.95	87.13	95.63	93.11	93.47
PGD	93.25	62.47	65.92	47.77	61.23

Table 6.13: Mnist4 Average Accuracy, ϵ = 10

Attack	CNN6	Resnet-8	Resnet-20	Resnet-32	Resnet-50
No Attack	99.44	99.07	99.13	99.30	99.11
FGSM	97.40	95.61	98.01	97.61	97.64
PGD	95.16	89.77	97.05	96.99	97.39

Table 6.14: Mnist10 Average Accuracy, ϵ = 10

6.4 Saliency Maps of Adversarial Examples

Saliency maps provide visualizations of the attention regions that highlight areas on an image that contribute most to the model's output. In the experiments, the saliency maps are utilized to examine how the model's attention regions change as the magnitude of the perturbation increases. In the medical datasets saliency visualisations this study shows that most of the time the larger resnet 50 model tends to focus its attention on portions of the image that are not necessarily related to the diagnosis of the disease. Whereas, networks of reduced complexity are focused mainly on the correct region of interest that actually contributes to disease diagnostic. For the natural image experiments, each of the models attention regions are in close vicinity to one another.

The saliency maps of medical and natural image DNNs are analyzed to gain insight on

the impact of model complexity for adversarial robustness of deep neural networks. The goal is to expound on how and why models of reduced complexity produce comparable performance to large complex state of the art networks while maintaining a greater degree of adversarial robustness. In this study, the saliency maps and decision boundary data distribution are visualized as model complexity is reduced.

6.4.1 Medical Image Saliency Maps

For the medical image saliency visualizations in Figure 6.4 models of reduced complexity have attention regions that are more concentrated on the regions of interest whereas larger more complicated networks have attention regions that are more sporadic. This may indicate that the networks of greater complexity are not truly learning the correct biological textures and thus it is easier to attack a model that learned incorrect features. The question is why did the more complex networks not learn the correct biological textures? Perhaps the complexity of the network is far too great to correctly learn such complex biological textures. Further investigation will be conducted to verify.

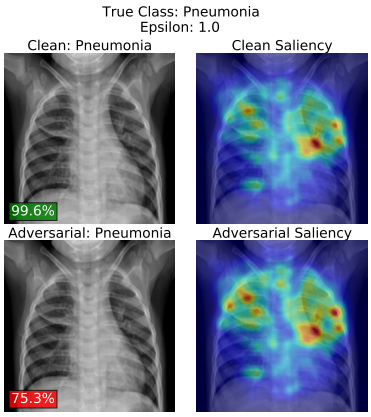
6.4.2 Cifar10 Saliency Maps

The cifar10 saliency visualizations in Figure 6.5 were fairly consistent with respect to the attention regions on clean data. Although, the adversarial saliency of the resnet50 networks were more sensitive to higher perturbations and result in greater change of the attention regions. This observation is in accordance with the performance curves that demonstrate that models of reduced complexity are more robust. The same sporadic attention region are not observed for resnet50 as seen in some of the medical image saliency visualizations. This could be due to the nature of the images, cifar10 does not contain images of biological texture and thus the network is able to focus the attention region on the regions that correctly contribute to the classification. Further investigation of texture analysis is required to reveal how the texture of the medical data contributes to the learning task

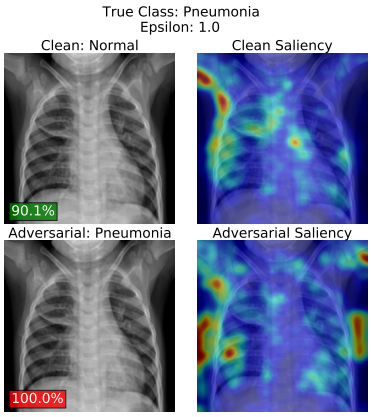
since this behavior is not observed in cifar10 saliency.

6.4.3 Mnist Saliency Maps

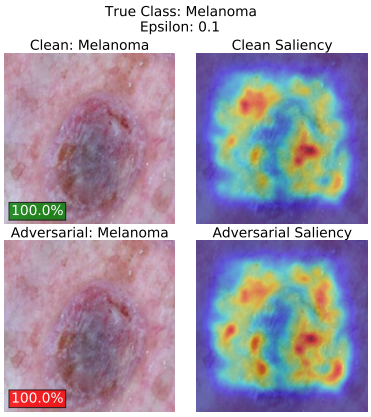
The mnist saliency visualizations in Figure 6.6 clearly focus the attention region on the correct object of interest for clean data. The magnitude of perturbation that most networks begin to experience performance degradation is $\epsilon = 10$. The saliency maps for this amount of perturbation make it obvious that the image background has been altered thus it is no longer imperceptible. One thing to note is that mnist is center focused and the object of interest is clear without any background noise so this may contribute to the level of robustness we observe for all models trained on mnist.



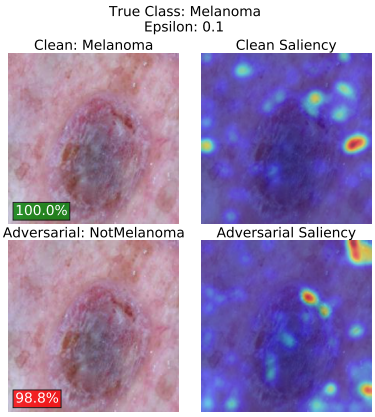
(a) Chest X-Ray, CNN5



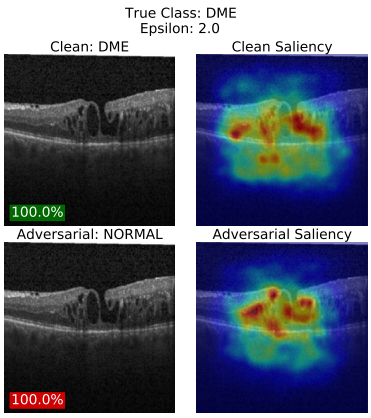
(b) Chest X-Ray, Resnet50



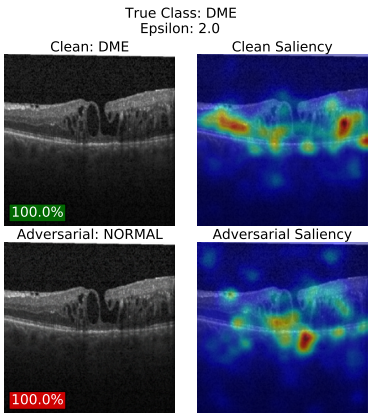
(c) Dermoscopy, CNN6



(d) Dermoscopy, Resnet50

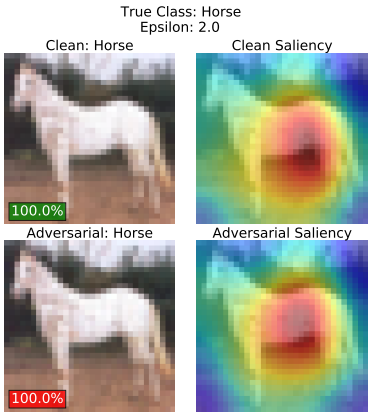


(e) OCT, CNN6

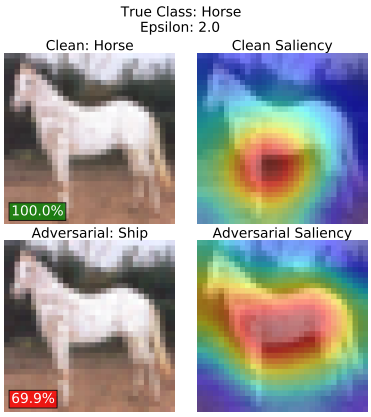


(f) OCT, Resnet50

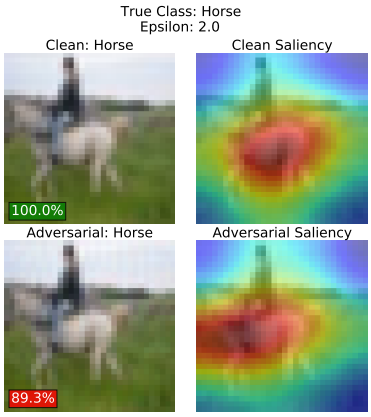
Figure 6.4: Mnist saliency maps for clean (column 1 & 3) and adversarial images (column 2 & 4) generated with CNN6 and Resnet50 respectively.



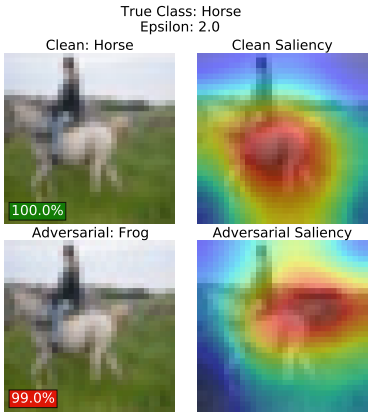
(a) Cifar2, CNN6



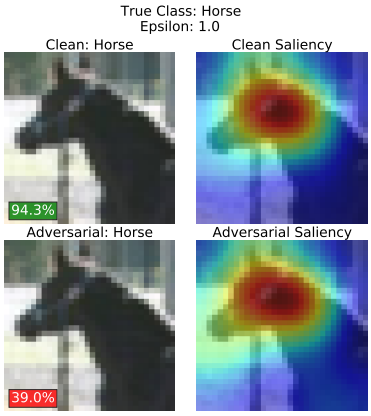
(b) Cifar2, Resnet50



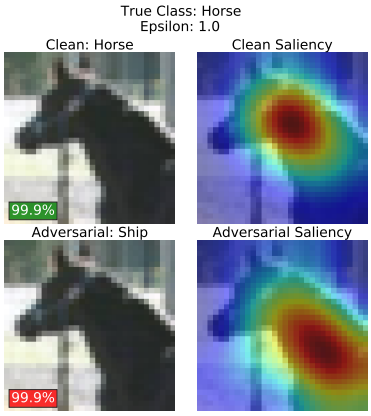
(c) Cifar4, CNN6



(d) Cifar4, Resnet50

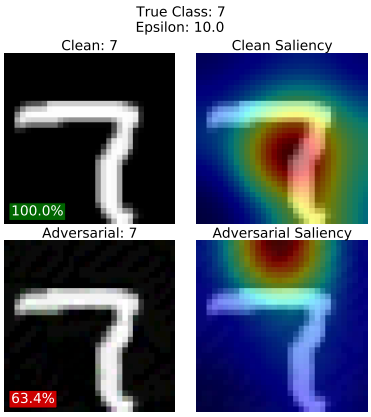


(e) Cifar10, CNN6

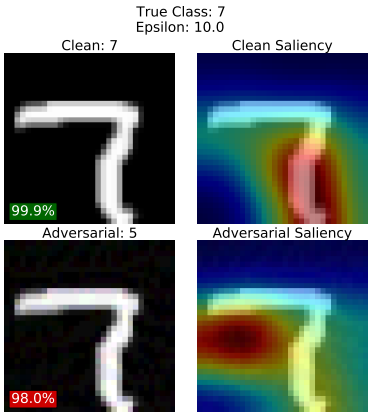


(f) Cifar10, Resnet50

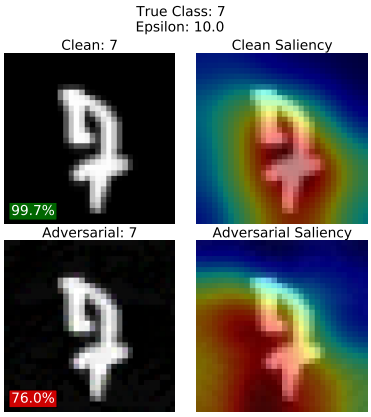
Figure 6.5: The cifar10 saliency maps for clean and adversarial images generated with CNN6 and Resnet50. Attention regions remain fairly consistent across networks on clean data.



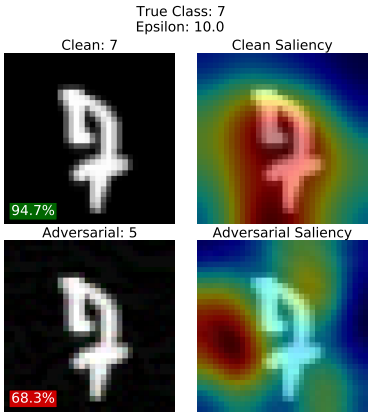
(a) Mnist2, CNN6



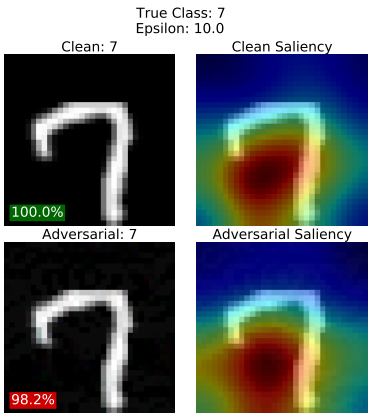
(b) Mnist2, Resnet50



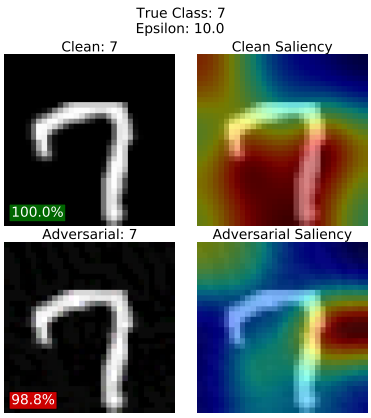
(c) Mnist4, CNN6



(d) Mnist4, Resnet50



(e) Mnist10, CNN6



(f) Mnist10, Resnet50

Figure 6.6: Mnist saliency maps for clean (column 1 & 3) and adversarial images (column 2 & 4) generated with CNN6 and Resnet50, respectively.

6.5 Decision Boundary Visualizations

For the decision boundary visualizations the data points gradually begin to cluster together as they migrate across the decision boundary, they begin to occupy a very small space. Although, models of higher complexity do not cluster as closely together at larger perturbations than models of reduced complexity. This may indicate that the larger search space could result in a more optimal perturbation than the models of reduced complexity, the latter does a better job at restricting the attacker from generating an optimal perturbation and could provide insight as to why models of greater complexity are less robust. Larger search spaces could potentially generate a stronger attack and with a smaller space one generates a weaker attack, this seems evident since the models of reduced complexity are more robust to attacks. To verify this we could generate adversarial examples for the higher complexity models to attack the lower complexity models and verify if the same level of robustness will still hold true.

The decision boundaries for each classification tasks are visualized as model complexity is reduced to understand how the data distribution is affected as the perturbation magnitude increases.

6.5.1 Visualization Procedure

In the decision boundary visualizations features are extracted from the second to last layer of a trained model and the features are used as input for t-distributed stochastic neighbor embedding (TSNE) [60]. The K-Nearest Neighbor classifier (KNN) [74] algorithm was utilized to fit a model on the combined data (train,test,validate,adversarial examples) while utilizing the predicted labels of the original neural network for each corresponding data point.

Specifically, adversarial examples were generated from a subset of the test dataset for a range of epsilon values and were then combined with the entire dataset. The previously trained

model was utilized for feature extraction on the combined dataset by obtaining the output of the layer prior to the classification layer. As a result, features of unperturbed and perturbed data are extracted, as they were all utilized to generate the TSNE projection which was performed to reduce the dimensionality of the image data and obtain a representation of the data distribution in 2D form. The purpose of this experiment was to replicate the decision boundary and data distribution to observe the behavior of adversarial examples on the decision boundary. The decision boundary visualizations were produced utilizing the mlxtend library [78]. The KNN classification algorithm was utilized to visualize the low dimensional data points produced by the TSNE projection, KNN calculates the euclidean distance between the data points and predicts a label based on how close the new data point is to data points that the model has stored.

Again, the goal was to investigate the behavior of adversarial examples on the decision boundary. Models of greater complexity allow smaller perturbations to influence the decision of the classifier. The decision boundaries of the models of reduced complexity demonstrate greater robustness. They require more perturbation for adversarial examples to migrate toward the opposite decision boundary. As the attack strength increases the data points become very closely clustered together, in other words the perturbations occupy a very small space as the magnitude of the attack increases.

6.5.2 Medical Data Decision Boundaries

The medical data decision boundary visualizations in Figure 6.7 display the activity on adversarial examples on the decision boundary. The magnitude of perturbation visualized were the points on the performance curve with the least amount of perturbation and the greatest margin between highest and lowest performing models. For most cases we observe that the resnet50 data points cluster much closer together along the decision boundary for a given epsilon. The corresponding data points for the standard CNN model do not exhibit much change in most cases. The OCT dataset exhibits slightly more clustering for the stand CNN than the chest x-ray and dermoscopy datasets.

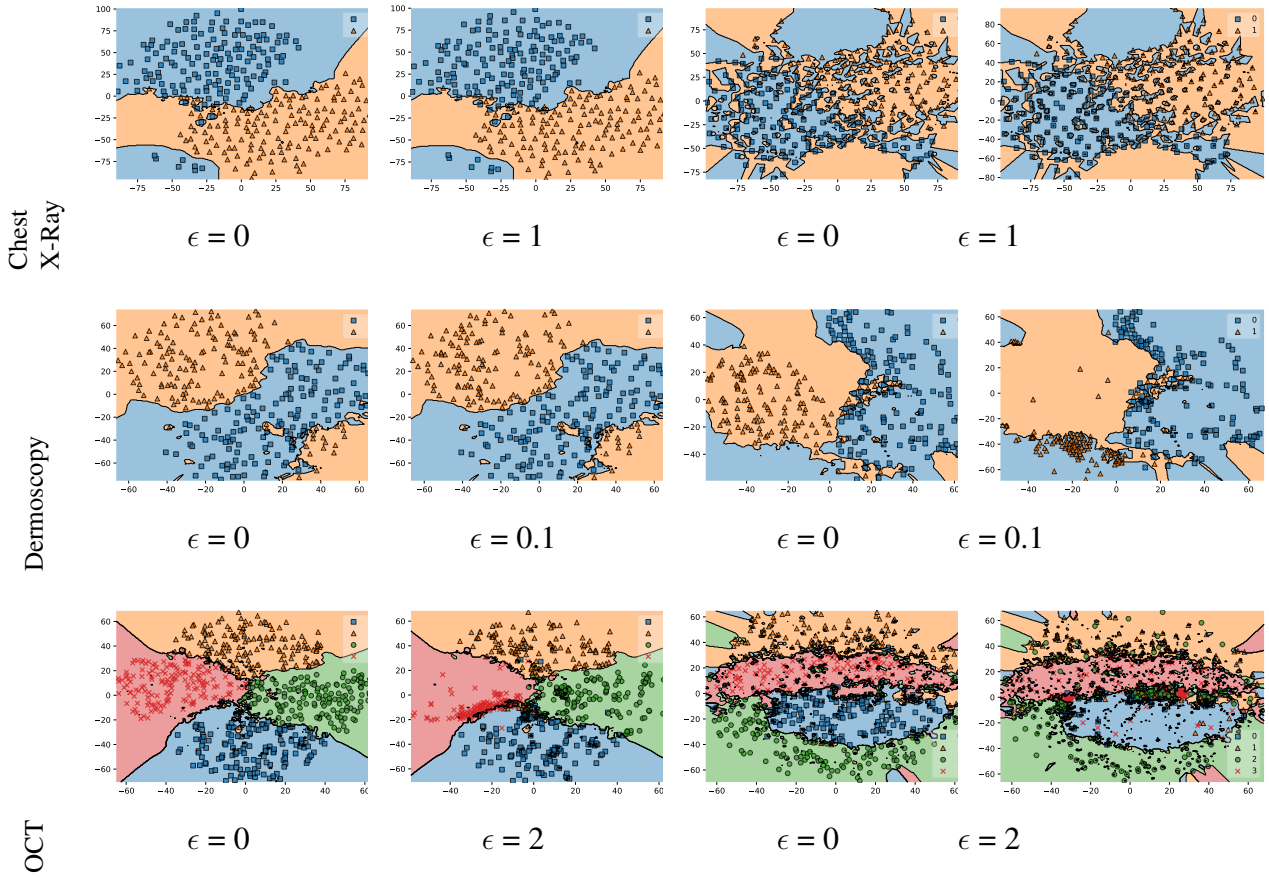


Figure 6.7: Adversarial examples on the decision boundary. Column 1 and 2 are the decision boundary visualizations for CBR-LargeT models before and after attacks, respectively. Column 3 and 4 are the decision boundary visualizations for Resnet-50 models before and after attacks, respectively.

This informs us that the number of classes impacts adversarial robustness of the network. The dermoscopy dataset visualization reveals that one class is more vulnerable to adversarial attacks for the resnet50 model. The chest x-ray visualization displays very little movement and clustering for both networks, although the resnet50 data points do slightly move into the opposing region boundary.

6.5.3 Cifar10 Decision Boundaries

The cifar10 data decision boundary visualizations in Figure 6.8 display the activity of adversarial examples on the decision boundary for the cifar10 dataset. Similar to the medical datasets, the magnitude of perturbation visualized were the points on the performance curve with the least amount of perturbation and the greatest margin between highest and lowest performing models. For most cases we observe that the resnet50 data points cluster much closer together along the decision boundary for a given epsilon. The corresponding data points for the standard CNN model do not exhibit much change in most cases. For cifar10 there is not much change as the number classes increased as was the case for the OCT medical dataset. Again, for cifar10 the larger more complex network begins to form a cluster that corresponds to the performance degradation seen in the performance curves. This may indicate that the larger search space may contribute to allowing the attacker to generate a more optimal perturbation that results in a more successful attack. The networks of reduced complexity may benefit from a smaller search space thus disallowing the attack to generate the most optimal perturbation. Further analysis on the search space could provide insight on the attackers ability to generate the optimal perturbation.

6.5.4 Mnist Decision Boundaries

The mnist data decision boundary visualizations in Figure 6.9 display the activity of adversarial examples on the decision boundary for the mnist dataset. Similar to the medical datasets, the magnitude of perturbation visualized were the points on the performance curve with the least amount of perturbation and the greatest margin between highest and lowest performing models. For most cases the data points cluster fairly close together along the decision boundary for a given epsilon. This behavior is expected since the performance curves exhibit consistent behavior across most networks. The mnist4 dataset does cluster more tightly for resnet50 which can also be attributed to larger search space allowing for a more optimal perturbation. Finding an optimal perturbation

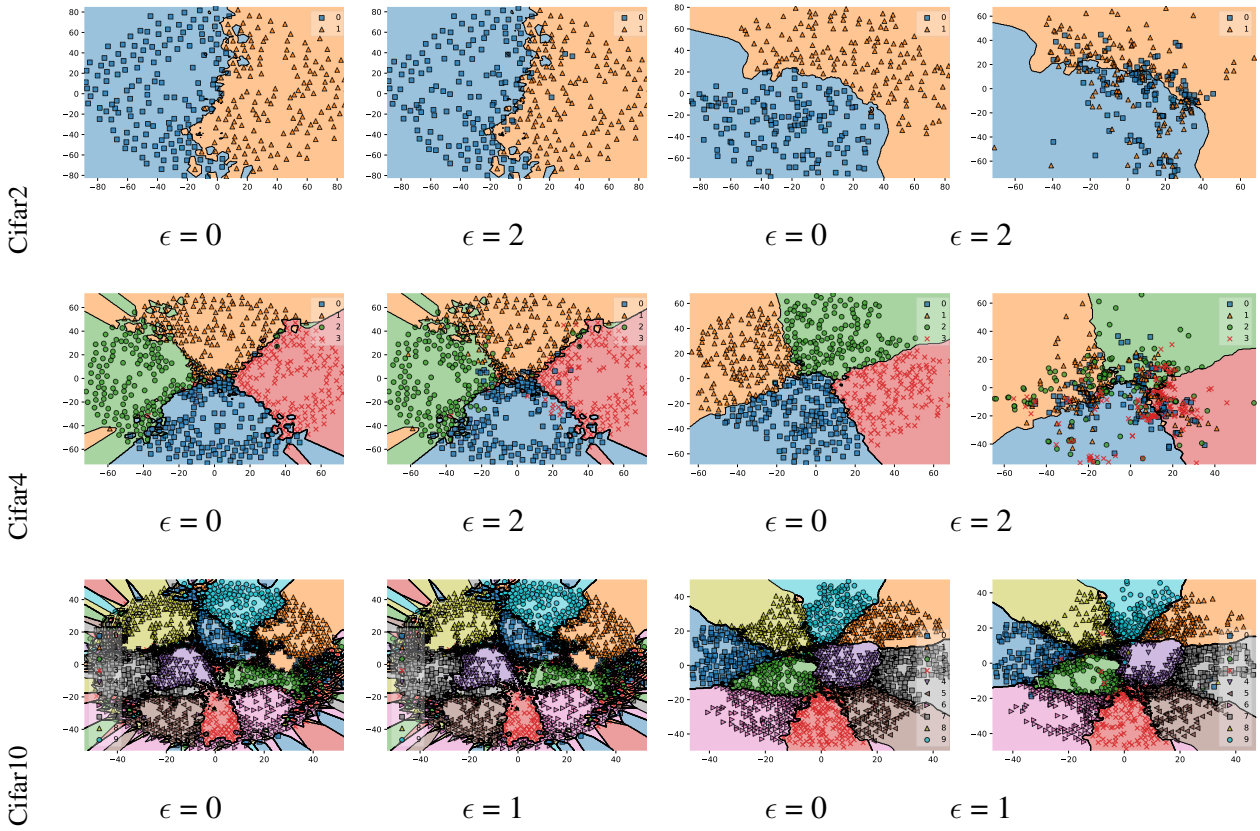


Figure 6.8: Adversarial examples on the decision boundary. Column 1 and 2 are the decision boundary visualizations for CNN6 models before and after attacks, respectively. Column 3 and 4 are the decision boundary visualizations for Resnet-50 models before and after attacks, respectively.

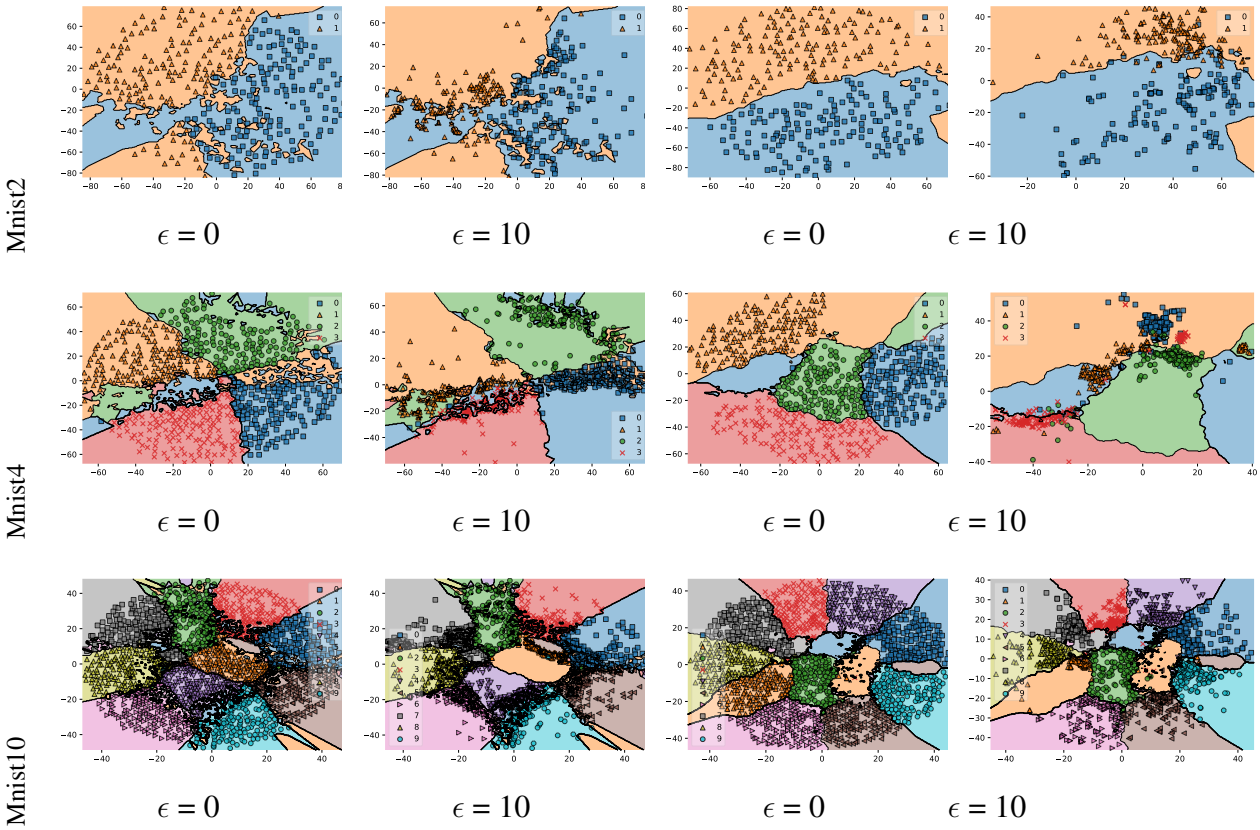


Figure 6.9: Adversarial examples on the decision boundary. Column 1 and 2 are the decision boundary visualizations for CNN6 models before and after attacks, respectively. Column 3 and 4 are the decision boundary visualizations for Resnet-50 models before and after attacks, respectively.

corresponds to a stronger attack which will result in higher performance degradation. Based on the performance curves there is an observable trend, as performance degrades on the curve the data points on the decision boundary become more tightly clustered. Also, for mnist4 the networks of reduced complexity may benefit from a smaller search space thus disallowing the attack to generate the most optimal perturbation. Further analysis on the search space could provide insight on the attackers ability to generate the optimal perturbation.

CHAPTER 7: CONCLUSION

In this study, learnable image transformation schemes were developed and evaluated using convolutional autoencoder and vision transformer to enhance privacy of deep learning models. An autoencoder-based image anonymization method was introduced for privacy enhance image classification. Additionally, an approach was developed for adversarially robust deep learning model selection.

7.1 Robustness Against Reconstruction Attacks

This study investigated two learnable image transformation schemes using convolutional autoencoder latent representation and vision transformer linear projection embeddings for privacy enhanced deep learning. The effectiveness of CAE and ViT image encoding schemes to preserve model utility was evaluated using classification accuracy and robustness to reconstruction attacks was evaluated using SSIM. The reconstruction attack methods used in this study consist of *Public Encoder Attack, Query Encoder Attack, Minimal Data Subset Attack and Cycle GAN Reconstruction Attack*. The image transformation schemes were demonstrated to be robust against reconstruction attacks. The results on Fashion Mnist, Cifar-10 and Chest X-ray datasets confirm that the investigated encoding schemes protect visual feature information of image data while preserving model utility.

7.2 Autoencoder-based Image Anonymization

This study developed an autoencoder-based image anonymization method using a standard convolutional autoencoder and multi-ouput resnet50 model to enhance the privacy of raw image data. The images were transformed into unrecognizable versions of the original input data. Highly

relevant feature information that is useful for classification was captured in the encoded images. Additionally, privacy is increased through the reduction of identity classification accuracy using the transformed images. In this work, it was demonstrated that the proposed method protects raw data features in the original images and enhance privacy of identity feature information while maintaining model utility with high attribute classification accuracy. In the experiments, the effectiveness of the image anonymization method was evaluated by measuring the reduction of attribute and identity classification accuracy. The experimental results confirm that the proposed method not only enables the maintaining of high image attribute classification accuracy but also enables the reduction of image identity classification accuracy.

7.3 Robustness Against Adversarial Attacks

In this study, the role of deep learning model complexity in adversarial robustness for image data was evaluated. It was demonstrated that standard trained medical image deep learning models of reduced complexity are more robust to adversarial attacks than large overly complex networks. This study showed that medical image deep learning models are more adversarially robust as model complexity decreases. The saliency map visualizations reveal that standard trained models of reduced complexity learn the features that contribute to the classification of disease better. The decision boundary visualizations show that larger overly complex networks result in data samples that are closer to the decision boundary in the projected space which increase the sensitivity of medical image deep learning models to input perturbations. The findings of this study provide guidance in deep learning model selection. Practitioners in the medical community should first evaluate the performance of a given set of deep learning model candidates on unperturbed medical images to realize networks of comparable performance and select the least complex model among the realized networks to produce the greatest robustness against adversarial attacks.

The adversarial robustness of deep neural networks trained on medical and natural images were evaluated, this study finds that models of reduced complexity produce greater adversarial

robustness than large complex state of the art networks. There is a rather uniform behavior across all most network for models of reduced complexity, that is, less perturbation is required to cause the model to misclassify. These networks demonstrate a higher degree of robustness than models trained using larger more complex architectures. The experiments verify that model complexity is crucial for adversarial robustness.

Consider a set of deep learning models that exhibit similar performances for a given task. These models are trained in the usual manner but are not trained to defend against adversarial attacks. This work demonstrates that, among those models, simpler models of reduced complexity show a greater level of robustness against adversarial attacks than larger models that often tend to be used in medical applications. On the other hand, this work also shows that once those models undergo adversarial training, the adversarial trained medical image deep learning models exhibit a greater degree of robustness than the standard trained models for all model complexities.

The above result has a significant practical relevance. When medical practitioners lack the expertise or resources to defend against adversarial attacks, the results of this study show that they should select the smallest of the models that exhibit adequate performance. Such a model would be naturally more robust to adversarial attacks than the larger models.

This concludes this manuscript. Following is the bibliography.

BIBLIOGRAPHY

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Sungtae An, Cao Xiao, Walter F Stewart, and Jimeng Sun. Longitudinal adversarial attack on electronic health records data. In *The World Wide Web Conference*, pages 2558–2564, 2019.

[3] Deepak Anand, Darshan Tank, Harshvardhan Tibrewal, and Amit Sethi. Self-supervision vs. transfer learning: Robust biomedical image analysis against adversarial attacks. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1159–1163. IEEE, 2020.

[4] Yoshinori Aono, Takuya Hayashi, Le Trieu Phong, and Lihua Wang. Scalable and secure logistic regression via homomorphic encryption. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, pages 142–144, 2016.

[5] Mikhail J Atallah, Konstantinos N Pantazopoulos, John R Rice, and Eugene E Spafford. Secure outsourcing of scientific computations. In *Advances in Computers*, volume 54, pages 215–272. Elsevier, 2002.

[6] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27:2654–2662, 2014.

[7] Eric Balkanski, Harrison Chase, Kojin Oshiba, Alexander Rilee, Yaron Singer, and Richard Wang. Adversarial attacks on binary image recognition systems, 2020.

[8] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Mesko. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1):1–8, 2020.

[9] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of Machine Learning and Systems*, 1:374–388, 2019.

[10] Charlotte Bonte and Frederik Vercauteren. Privacy-preserving logistic regression training. *BMC medical genomics*, 11(4):13–21, 2018.

[11] Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Shuang Song, Abhradeep Thakurta, and Florian Tramer. Is private learning possible with instance encoding?, 2020.

[12] Nicholas Carlini, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Florian Tramer. Neuracrypt is not private, 2021.

[13] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey, 2018.

[14] Melissa Chase, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, and Peter Rindal. Private collaborative neural network learning. *Cryptology ePrint Archive*, 2017.

[15] Zhenfei Chen, Tianqing Zhu, Ping Xiong, Chenguang Wang, and Wei Ren. Privacy preservation for image data: a gan-based method. *International Journal of Intelligent Systems*, 36(4):1668–1685, 2021.

[16] François Chollet et al. Keras. `https://keras.io`, 2015.

[17] Jack L. H. Crawford, Craig Gentry, Shai Halevi, Daniel Platt, and Victor Shoup. Doing real work with fhe: The case of logistic regression. Cryptology ePrint Archive, Paper 2018/202, 2018. https://eprint.iacr.org/2018/202.

[18] Ekin D Cubuk, Barret Zoph, Samuel S Schoenholz, and Quoc V Le. Intriguing properties of adversarial examples. *arXiv preprint arXiv:1711.02846*, 2017.

[19] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[20] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[23] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[24] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[25] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[26] Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane, and Andrew L. Beam. Adversarial attacks against medical deep learning systems, 2018.

[27] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[28] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

[29] Thore Graepel, Kristin Lauter, and Michael Naehrig. Ml confidential: Machine learning on encrypted data. In *International Conference on Information Security and Cryptology*, pages 1–21. Springer, 2012.

[30] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr DollÃ¡r, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[34] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[35] Hokuto Hirano, Akinori Minagi, and Kazuhiro Takemoto. Universal adversarial attacks on deep neural networks for medical image classification. *BMC medical imaging*, 21(1):1–13, 2021.

[36] Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. InstaHide: Instance-hiding schemes for private distributed learning. In Hal DaumÃ© III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4507–4518. PMLR, 13–18 Jul 2020.

[37] Zonghao Huang, Rui Hu, Yuanxiong Guo, Eric Chan-Tin, and Yanmin Gong. Dp-admm: Admm-based distributed learning with differential privacy. *IEEE Transactions on Information Forensics and Security*, 15:1002–1012, 2019.

[38] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features, 2019.

[39] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590â597, Jul 2019.

[40] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

[41] ISSN International Centre. The issn register, 2006.

[42] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1895–1912, 2019.

[43] X Jones. Zeolites and synthetic mechanisms. In Y Smith, editor, *Proceedings of the First National Conference on Porous Sieves: 27-30 June 1996; Baltimore*, pages 16–27, 1996.

[44] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[45] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y.L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122 – 1131.e9, 2018.

[46] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[47] Miran Kim, Yongsoo Song, Shuang Wang, Yuhou Xia, Xiaoqian Jiang, et al. Secure logistic regression based on homomorphic encryption: Design and evaluation. *JMIR medical informatics*, 6(2):e8805, 2018.

[48] R Kohavi. *Wrappers for performance enhancement and obvious decision graphs*. PhD thesis, Stanford University, Computer Science Department, 1995.

[49] Jakub KoneÄnÃœ, H. Brendan McMahan, Felix X. Yu, Peter RichtÃ¡rik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency, 2016.

[50] E V Koonin, S F Altschul, and P Bork. Brca1 protein products: functional motifs. *Nat. Genet.*, 13:266–267, 1996.

[51] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

[52] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2016.

[53] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale, 2016.

[54] Y. LECUN. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*.

[55] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[56] Tiancheng Li and Ninghui Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–526, 2009.

[57] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[58] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, page 107332, May 2020.

[59] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021.

[60] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[61] K Madono, M Tanaka, M Onishi, and T Ogawa. An adversarial attack to learnable encrypted images. In *22nd IEICE Symposium on Image Recognition and Understanding*, 2019.

[62] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017.

[63] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*, pages 49–58, 2019.

[64] L Margulis. *Origin of Eukaryotic Cells*. Yale University Press, New Haven, 1970.

[65] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016.

[66] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[67] Payman Mohassel and Yupeng Zhang. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 19–38, 2017.

[68] Karthik Nandakumar, Nalini Ratha, Sharath Pankanti, and Shai Halevi. Towards deep neural network training on encrypted data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[69] Valeria Nikolaenko, Udi Weinsberg, Stratis Ioannidis, Marc Joye, Dan Boneh, and Nina Taft. Privacy-preserving ridge regression on hundreds of millions of records. In *2013 IEEE Symposium on Security and Privacy*, pages 334–348, 2013.

[70] Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

[71] Kai Packhäuser, Sebastian Gündel, Nicolas Münster, Christopher Syben, Vincent Christlein, and Andreas Maier. Is medical chest x-ray data anonymous? *arXiv preprint arXiv:2103.08562*, 2021.

[72] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.

[73] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[75] Bijayalaxmi Purohit and Pawan Prakash Singh. Data leakage analysis on cloud computing. *International Journal of Engineering Research and Applications*, 3(3):1311–1316, 2013.

[76] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging, 2019.

[77] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017.

[78] Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to pythonâs scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018.

[79] Vibhor Rastogi, Dan Suciu, and Sungho Hong. The boundary between privacy and utility in data publishing. In *Proceedings of the 33rd international conference on Very large data bases*, pages 531–542, 2007.

[80] Md. Fazle Rasul, Nahin Kumar Dey, and M.M.A. Hashem. A comparative study of neural network architectures for lesion segmentation and melanoma detection. 06 2020.

[81] Mathilde Raynal, Radhakrishna Achanta, and Mathias Humbert. Image obfuscation for privacy-preserving machine learning, 2020.

[82] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211â252, Apr 2015.

[83] Alexander Scarlat. 'dermoscopic pigmented skin lesions from ham10k', 2019. `https://www.kaggle.com/drscarlat/melanoma"`, (Accessed: 02/5/2020).

[84] E Schnepf. From prey via endosymbiont to plastids: comparative studies in dinoflagellates. In R A Lewin, editor, *Origins of Plastids*, pages 53–76. Chapman and Hall, New York, 2nd edition, 1993.

[85] Soumitra Sengupta, Neil S. Calman, and George Hripcsak. A Model for Expanded Public Health Reporting in the Context of HIPAA. *Journal of the American Medical Informatics Association*, 15(5):569–574, 09 2008.

[86] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017. PMID: 28301734.

[87] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

[88] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[89] Warit Sirichotedumrong, Yuma Kinoshita, and Hitoshi Kiya. On the security of pixel-based image encryption for privacy-preserving deep neural networks. *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, pages 121–124, 2019.

[90] Warit Sirichotedumrong, Yuma Kinoshita, and Hitoshi Kiya. Pixel-based image encryption without key management for privacy-preserving deep neural networks. *IEEE Access*, 7:177844–177855, 2019.

[91] Warit Sirichotedumrong and Hitoshi Kiya. A gan-based image transformation scheme for privacy-preserving deep neural networks, 2020.

[92] Warit Sirichotedumrong, Takahiro Maekawa, Yuma Kinoshita, and Hitoshi Kiya. Privacy-preserving deep neural networks with pixel-based image encryption considering data aug-

mentation in the encrypted domain. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 674–678, 2019.

[93] Chang Song, Hsin-Pai Cheng, Huanrui Yang, Sicheng Li, Chunpeng Wu, Qing Wu, and Hai Li. Adversarial attack: A new threat to smart devices and how to defend it. *IEEE Consumer Electronics Magazine*, 9(4):49–55, 2020.

[94] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy? â a comprehensive study on the robustness of 18 deep image classification models. *Lecture Notes in Computer Science*, page 644â661, 2018.

[95] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2013.

[96] Masayuki Tanaka. Learnable image encryption. In *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2, 2018.

[97] Florian TramÃšr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses, 2020.

[98] Philipp Tschandl. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. 2018.

[99] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2019.

[100] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[101] Sameer Wagh, Divya Gupta, and Nishanth Chandran. Securenn: 3-party secure computation for neural network training. *Proc. Priv. Enhancing Technol.*, 2019(3):26–49, 2019.

[102] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 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)*, Jul 2017.

[103] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. *arXiv preprint arXiv:2002.05990*, 2020.

[104] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):209–226, 2016.

[105] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[106] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[107] Adam Yala, Homa Esfahanizadeh, Rafael G. L. D' Oliveira, Ken R. Duffy, Manya Ghobadi, Tommi S. Jaakkola, Vinod Vaikuntanathan, Regina Barzilay, and Muriel Medard. Neuracrypt: Hiding private health data via random neural networks for public training, 2021.

[108] Andrew C Yao. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*, pages 160–164. IEEE, 1982.

[109] Shaokai Ye, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, Yanzhi Wang, and Xue Lin. Adversarial robustness vs. model compression, or both? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 111–120, 2019.

[110] Gu Yonghao and Wu Weiming. A quantifying method for trade-off between privacy and utility. In *IET International Conference on Information and Communications Technologies (IETICT 2013)*, pages 270–273. IET, 2013.

[111] Xingliang Yuan, Xinyu Wang, Cong Wang, Anna Squicciarini, and Kui Ren. Enabling privacy-preserving image-centric social discovery. In *Proceedings of the 2014 IEEE 34th International Conference on Distributed Computing Systems*, ICDCS '14, page 198â207, USA, 2014. IEEE Computer Society.

[112] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[113] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017.

VITA

David Rodriguez received his B.S. degree in Electrical Engineering from the University of Texas at San Antonio in 2017 and his M.S. degree in Computer Engineering from the University of Texas at San Antonio in 2018. He is currently pursuing his Ph.D in Electrical Engineering at the University of Texas at San Antonio since 2019. In 2018 he worked as a graduate research assistant in blockchain technology utilizing hyper ledger fabric framework. His current research is in evaluating adversarial robustness of DNN models on medical imaging. Additionally, his current research includes leveraging an autoencoder-based images transformation scheme for privacy enhanced deep neural networks. His research interests include deep neural networks, privacy, security, image anonymization and adversarial machine learning on medical imaging.