SECURING ADVERSARIAL MACHINE LEARNING IN MEDICAL IMAGING APPLICATIONS

GAURANG M PATEL

Bachelor of Technology in Electronics and Communications Engineering

Dharmsinh Desai University

May 2015

submitted in partial fulfillment of requirements for the degree

MASTER OF COMPUTER AND INFORMATION SCIENCE

at the

CLEVELAND STATE UNIVERSITY

August 2023

We hereby approve this thesis for

GAURANG M PATEL

Candidate for the Master of Science in Computer Science degree for the

Department of Computer Science

and the CLEVELAND STATE UNIVERSITY'S

College of Graduate Studies by

Thesis Committee Chairperson, Dr. Sathish Kumar, Ph.D.

Department & Date

Thesis Committee Member, Dr. Janche Sang, Ph.D.

Department & Date

Thesis Committee Member, Hongkai Yu, Ph.D.

Department & Date

Student's Date of Defense: August 3, 2023

DEDICATION

TO MY FAMILY

ACKNOWLEDGEMENT

I would like to thank Dr. Sathish Kumar for the opportunity, his continued support and guidance throughout my research experiment. He has helped me learn machine learning and convolutional neural networks better. I would like to thank Dr. Vivek Chaturvedi for his guidance and meeting with me whenever I needed help. I would also like to thank Dr. Janche Sang and Dr. Hongkai Yu for serving on my thesis committee.

Finally, I would like to thank my family, this would not have been possible without their support and belief in me.

SECURING ADVERSARIAL MACHINE LEARNING IN MEDICAL IMAGING APPLICATIONS

GAURANG M PATEL

ABSTRACT

Deep learning has revolutionized several fields including the medical image processing in the past decade. Convolutional Neural Networks can now perform many image processing tasks better than humans. As a result, Convolution Neural Networks (CNNs) are increasingly used in the automation of diagnosis of life-threatening diseases. CNNs perform complex image classification tasks with greater accuracy and output quality. However, recent discovery of adversarial attacks raises a significant threat against safety and accuracy of the CNNs. CNNs are vulnerable to perturbations in the input image that are imperceptible to human eyes, which leads to misclassification of the model output. This research work proposes a novel Super Resolution Generative Adversarial Network-based approach to improve classification robustness of CNN against adversarial attacks using MRI dataset as an example. Robustness of proposed novel network model is compared with existing state of the art models in the field. The experiment results demonstrate that proposed approach improves CNN model robustness by 95% against adversarial attacks when compared to state-of-the-art approaches such as context-aware-models and conventional CNN.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND BACKGROUND

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that includes algorithms which build a model based on sample input data and then this model will be used to predict the output using unseen test data. ML lifecycle has three stages: training, evaluation, and deployment. First stage is training where ML model learn from the training data. In the evaluation stage, model is evaluated for its performance on a never-before-seen test dataset. If performance satisfies the requirement, model is deployed for the real life usage.

ML includes various learning algorithms such as linear regression, logistic regression, support vector machine, decision tree, deep learning etc. Out of these algorithms, Deep Learning (DL) has gained popularity in past decade thanks to the state-of-the-art performance and availability of affordable GPUs. DL has been applied to many complex problems such as financial market prediction [1], image generation from text [2] and natural language processing [3]. At the heart of DL is the neural network (NN), a network made up of artificial neurons. Deep neural networks contain an input layer, one or more hidden layer and a final output layer (Figure 1).

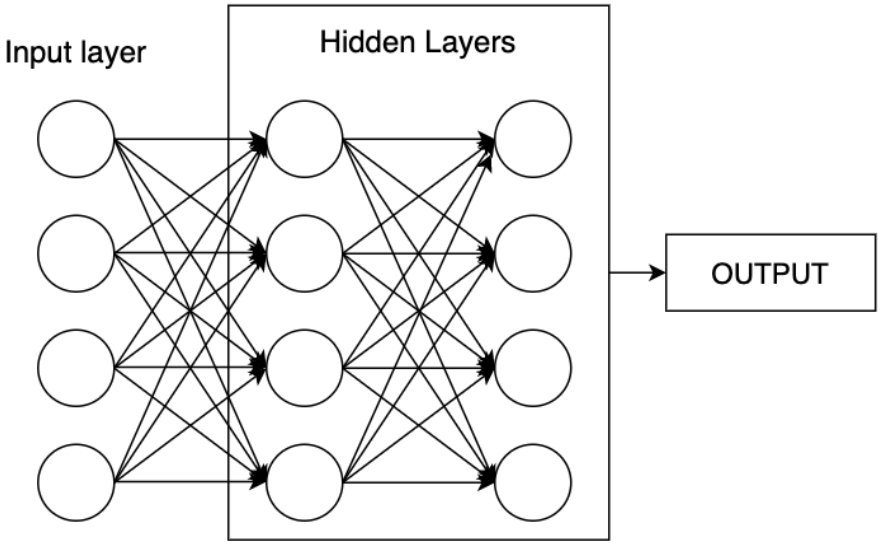

Figure 1 A Neural Network

For Computer Vision tasks, Convolutional Neural Networks are state-of-the-art algorithms. CNNs are NNs with additional convolutional and pooling layers to help with feature extractions from images. CNNs have helped rapidly advance the fields such as self-driving cars [4], handwritten character recognition [5] and image classification [6]. CNN architectures have evolved quickly with advances in capability of GPUs and network architectures, achieving human level performance in tasks such as image classification [7].

1.1 Machine Learning in Medical Image Processing

Medical image processing is one of the areas where Deep Learning can help solve many problems which were previously unsolved or were not scalable due to complexities. Scalability of neural networks can help us deal with future shortage of healthcare professionals [8]. Once trained, neural networks-based ML systems can be deployed as many times as needed. In recent years, Deep Learning has been applied to many medical imaging problems such as diagnosis of cervical cancer [9], detection of diabetic retinopathy [10] and semantic segmentation of organs , risk for prostate cancer radiotherapy [11]. The research community has utilized CNNs on various types of cancers that can be diagnosed

from images. Many of the work in medical image processing use the neural networks trained on the natural images and train them further to work on medical image processing, which saves computing resources as well as time. These neural networks achieve human level accuracy even in medical image classification [12]. In 2018, FDA approved marketing of artificial intelligence (AI)-based device to detect diabetic retinopathy which can be used without the clinician [13].

Despite the wide range of applications of deep learning and state-of-the-art performance, it has been shown that these ML algorithms are susceptible to adversarial attacks. Various approaches are being proposed on how to create adversarial input range through simple gradient-based attacks [14] to complex and computation heavy attacks presented as optimization problems [15]. Various works have tried to justify existence of adversarial examples. Goodfellow et al. [14] use linearity hypothesis, Szegedy et al. [16] first explained it using non-linearity hypothesis while Tanay et al. [17] create adversarial examples using boundary tilting hypothesis.

We discussed how deep learning can help revolutionize medical image processing. If we were to consider adversarial attacks, it becomes clear that they open door to new fraud and harm across healthcare industry. In fact, Ma et al. [18] show that it is easier to create adversarial inputs for the medical images than it is to create adversarial inputs for natural images. Adversarial attacks affect all areas of healthcare industry where deep learning is applied which include diagnose, treatment, insurance, billing and many more. Considering that US healthcare industry spending was $4.1 trillion in 2020 and is projected to reach $6.2 trillion in 2028 [19], it presents lucrative opportunities for bad actors. There have been many defenses proposed to deal with adversarial attacks, but no single solution

is effective against all attacks. Given that healthcare systems deal with human lives and are difficult to update, they need to be much more robust. This research work aims to address the problem of adversarial attacks in medical imaging systems that use brain MRIs.

1.2 Adversarial Machine Learning

This sections provides necessary background and vocabulary for the adversarial machine learning. Definition and categories of adversarial input as well as threat models based on adversarial attacks and adversarial goals are discussed below.

1.2.1 Adversarial Input

As per the definition of adversarial example in [20], let us consider the neural network to be a function f, which takes input x and maps it to its label y, which we will denote as $f(x) = y$. Adversarial input x' can be created by adding perturbation Δx to the original input x such that $x' = x + \Delta x$ and $f(x) \neq f(x')$. Taxonomy of adversarial images can have three different axes: (i) perturbation scope, (ii) perturbation visibility and (iii) perturbation measurement.

Perturbations can either be individually scoped or universally scoped. Individual scoped perturbations are more prevalent in research. Universally scoped perturbations are applied on all available data or a batch of data. Since universally scoped perturbations are image-agnostic, they make it easier to create adversarial input at scale. Perturbation visibility relates to the visibility of perturbations to humans and to deep learning models. Perturbation is measured using L_p norms. In this experiment, we use p-norms L_0 and L_∞ to measure perturbation. p norm can be defined as:

$$L_p = \sqrt[p]{\sum |x - x'|^p} \quad (1)$$

Here p is the dimension used to measure the norm.

1.2.2 Threat models

Threat models categorically define how vulnerabilities are exploited by attacker that affects the victim. This view of the system from the security perspective helps us in implementing defense against possible threats. Here we discuss threat models defined by [21] from the perspective of adversarial goals and attacker's knowledge.

1.2.2.1 Adversarial Goals

Main intent of the adversarial attacks is to cause misclassification of the input with added perturbation. There are the adversarial goals that affect the functionality of deep learning systems:

1. **Confidence reduction**: In this type of attack, perturbed input causes the neural network to reduce its confidence in predicted class compared to the legitimate input.
2. **Misclassification**: Goal here is to change the prediction of the neural network to entirely different class.
3. **Targeted misclassification**: In this type of attack, attacker wants to change the prediction of the model to a desired class.
4. **Source/Target misclassification**: The attacker tries to create a perturbed image in such a way that the neural network changes its prediction from the fixed source class of input to the desired target class.

It is important to mention that the difficulty of creating the desired perturbed image increase as we move from top to bottom on the above list. Hence, Source/Target misclassification is the most difficult to achieve. In this work, white box attacks aim for misclassification and black box attacks aim for targeted misclassification.

1.2.2.2 Adversarial knowledge/Adversarial Capabilities

Complexity and effectiveness of the attack greatly depends on the knowledge an attacker has about the system. The attacks can be classified based on attacker's knowledge of the system as shown below.

White Box Attacks:

In this category of attack, the attacker possesses knowledge about the entire or a part of the system, as defined below:

1. **Architecture and Training Data**: When attacker has access to network architecture as well as training data, they can poison the training data and use knowledge of architecture to their advantage. Gradient-based attacks also fall into this category.
2. **Architecture**: In this case, the attacker will generate synthetic training data and use that data with architecture knowledge to perform the attack.
3. **Training Data**: If attackers only have access to training data, they can use predictions from original model and train a surrogate model. Adversarial attacks are transferable between different architecture of ML models which makes this and next type of attacks possible and easy.

Black Box attacks:

4. **Oracle**: When attacker has access to neither network architecture nor training data, system is treated as the Oracle and synthetic data is used to make prediction. This is a black box attack. Attacker can then train surrogate model and perform attack on that model. This attack is hardest among the list but if

attacker can create a surrogate model, they can take advantage of transferability property.

1.3 Contributions

Age detection from patient MRI is an important problem in medical image processing [22]. It can be helpful in determining person's age when birth record is not available. CNNs can also predict chronological age of a person which helps in identifying deviations from healthy aging [23]. This work proposes a novel approach using a hybrid SRGAN and CNN model to predict age of a person from their brain MRI. This approach results in a system that is considerably more robust than existing systems. Adversarial attacks iteratively introduce minimal perturbations to the input image, which move prediction out of the original class boundary. Proposed approach makes use of SRGAN's property of recovering original image from low resolution image to protect against adversarial attacks. While recovering high resolution image, SRGAN also learns original feature space of the data. This characteristic will help remove the effect of small perturbations on each iteration and the generated image will be much closer to original input. This results in a highly robust system that can protect against white box as well as black box attacks. Contributions of this work can be summarized below:

- Proposed approach introduces a combination of SRGAN and CNN to predict age of the person from their brain MRI and evaluate their performance.
- We Compare proposed SRGAN-CNN hybrid approach against robustness of different CNN architectures including state-of-the-art Context-aware model against adversarial examples using white box and black box adversarial attacks.

- We demonstrate that our novel approach of using SRGAN with CNN is more robust against adversarial attacks than existing approaches.

1.4 Organization of The Thesis

The rest of the thesis is organized as follows: Section 2 discusses related research work and their limitations in the area of adversarial machine learning specifically in the medical image processing systems context. Section 3 states the research objective of this work. Section 4 explains the methodology of the experiment: datasets used, preprocessing of the input MRIs, network models used in the experiments, white box, and black box attacks. At the end of section 4, proposed new approach to protect from the adversarial machine learning attacks is discussed. Section 5 discusses results of the experiment and comparison between network models used in the experiment. Section 6 includes lessons learned during implementation of this experiment. In section 7 future work recommendations are discussed and finally Section 8 concludes this research work.

CHAPTER II

LITERATURE REVIEW

Unlike natural images, medical images have different modalities such as MRIs, X-ray, and CT-scan. Each of these modalities are captured with different devices and have different formats. As a result, implementation of preprocessing steps, neural networks structures, adversarial attacks vary across modalities. In this section, some of the works published in the area of adversarial attacks and detection in medical imaging machine learning systems are discussed. This provides a context for current research environment as well as an idea of a wide range of applications of ML in medical imaging. Table 1 lists summary of related works that implement adversarial attacks, their detection and defense against the attacks. These works are discussed in section 2.1, 2.2 and 2.3.

No.	Study	Models	Attack Algorithm	Work Type	Image Modality, Work Type, Results and Datasets
1	[24]	SegNet, U-Net, DenseNet	FGSM, DeepFool, SMA	Adversarial attacks	**Image Modality:** MRI **Datasets:** OASIS dataset **Task:** Brain segmentation **Results:** - Accuracy drops to 37%
2	[25]	U-Net	Custom algorithm for universal perturbation	Adversarial attacks	**Image Modality**: MRI **Datasets**: MICCAI BraTS 2019 **Task:** Brain segmentation **Results:** - Dice score drops to 0.31

3	[26]	ResNet-50, U-Net	FGSM, PGD, DeepFool, DAG, SMIA	Adversarial attacks	**Image Modality**: Fundoscopy, Endoscopy, CT-scan **Datasets**: Kaggle DR, APTOS-2019, EAD 2019, Kaggle COVID-19 CT scans **Task:** Diabetic retinopathy detection, Endoscopy artifact detection, Infected lung region segmentation **Results:** - SMIA attack is more successful in reducing accuracy compared to existing attacks
4	[27]	ResNet-50	PGD, Patch based attacks	Adversarial attacks	**Image Modality**: X-ray, Dermoscopy, Fundoscopy **Datasets**: Kaggle DR, ChestX-ray14, ISIC **Task:** Diabetic retinopathy detection, Lung disease classification, Malignant melanoma detection **Results:** - Accuracy and AUROC drops significantly
5	[18]	ResNet-50	FGSM, BIM, PGD	Adversarial attacks and detection	**Image Modality**: X-ray, Dermoscopy, Fundoscopy **Datasets**: Kaggle DR, ChestX-ray8, ISIC **Task:** Diabetic retinopathy detection, Lung disease classification, Malignant melanoma detection **Results:** - Accuracy drops to 0 against attacks. - AUC score of 100% for detection
6	[28]	DenseNet-121, ResNet- 50	FGSM, BIM, PGD, MIM	Adversarial attacks and detection	**Image Modality**: X-ray **Datasets**: ChestX-ray14 **Task:** Lung disease classification **Results:** - Attacks reduce F1 Score to 0.5 - Detection helps recover F1 score to ~0.9
7	[29]	Cascaded CNN	FGSM	Adversarial attacks and defense	**Image Modality**: MRI **Datasets**: MRBrainS18 dataset **Task:** MRI segmentation **Results:** - Defense improves the dice score to 83.12%

8	[30]	FNAF U-Net, FNAF I-RIM	FNAF	Adversarial attacks and defense	**Image Modality**: MRI **Datasets**: fastMRI knee dataset **Task:** MRI reconstruction **Results:** - Successful attack on 92% of the dataset - Reconstruction improved with adversarial training (defense)

Table 1 Related Works in Adversarial ML in medical imaging systems

2.1 Related works that implement Adversarial Attacks in Medical Imaging Applications

Paschali et al. [24] propose a new approach which uses adversarial examples to measure robustness of the model. They argue that current techniques of model evaluation give more importance to over-fitting while not paying enough attention to model sensitivity to input variations. They craft adversarial attack against classification models (Inception V3 [6], Inception V4 [31] and MobileNet [32]) and segmentation models SegNet (SN) [33], U-Net (UN) [34] and DenseNet (DN) [35]. To show the effect of adversarial examples against classification models Fast Gradient Sign Method (FGSM) [14], DeepFool (DF) [36] and Saliency Map Attacks (SMA) [37] are used. For segmentation models, Dense Adversarial Generation (DAG) [38] method is used with incorrect segmentation mask to create adversarial samples. Results show that adversarial samples are consistently misclassified compared to samples perturbed with gaussian noise. Out of all the models tested for classification, Inception V4 was more robust against DF and SMA, Inception V3 was more robust against FGSM. Segmentation experiments showed that DN was more robust against adversarial attacks as well as noise. The authors conclude that proposed approach accounts for generalizability as well as robustness of the models.

Cheng et al. [25] perform adversarial attacks on U-Net used for brain tumor segmentation. Brain segmentation models use all four modalities of the brain MRI images during training and hence adversarial samples are created for all modalities of brain MRI: T1, T2, FlAIR (Fluid-Attenuated Inversion Recovery), and contrast-enhanced T1. They use MICCAI BraTS 2019 dataset which provides MRIs from different MRI scanners and institutions. Adversarial noise is generated using two components: max norm vector ($vec1$) and l_2 norm vector ($vec2$). These two vectors are generated randomly using gaussian distribution. Total perturbation can be controlled by two hyperparameters: ϵ and rad. Perturbation is calculated using by below equation 1:

$$pert = \epsilon * sgn(vec1) + rad * vec2 \quad (2)$$

Their finding suggests that when the single modality of the dataset is attacked, rest of the modality help protect the network. If all modalities are attacked, performance of network reduces significantly.

Qi et al. [26] propose a new adversarial attack method named Stabilized Medical Image Attack (SMIA). SMIA uses an objective function containing a deviation loss term and a stabilized loss term. Perturbations are generated iteratively by taking partial derivative of the objective function with respect to the input image. SMIA attack is implemented against ResNet-50 trained on fundus images for diabetic retinopathy detection and U-Net trained on CT scans of lungs to segment infected lung regions. Comparison between FGSM, PGD, DeepFool, DAG and SMIA shows that SMIA was successful in reducing accuracy of networks to their lowest.

Finlayson et al [27] use a RestNet-50 model pretrained on ImageNet dataset. This model is then fine-tuned on medical imaging datasets resulting in one model for every task.

They implement human imperceptible white box and black box PGD and patch-based attack on the model. PGD is an iterative version of FGSM attack. Patch-based attack is implemented by finding a patch for the entire training dataset which makes the adversarial patch universal. White box version of the attacks is performed directly on the original model. Black box attacks are implemented by training a surrogate model, crafting adversarial inputs for the surrogate model, and then using the same adversarial inputs on the original model. This approach takes advantage of transferability property of the adversarial attacks discussed in section 1.2.2.2. The results show that for lung disease classification, white box PGD and white box patch-based attacks reduce accuracy of the model from 94.9% to 0%. Rest of the patch-based and PGD attacks also bring accuracy to < 15%. For Fundoscopy and Dermoscopy datasets, these attacks are successful in lowering accuracy and reducing AUROC to 0.

2.2 Related works that detect Adversarial Attacks in Medical Imaging Applications

Work presented in [18], aims to answer the question "Is crafting adversarial attacks on medical images is as easy as attacks on natural images?". Their findings suggest that adversarial attacks are much more successful on medical images. They argue that this is due to complex biological structures present in medical images that result in high gradient regions. Also, neural networks designed for natural image processing are not suitable for medical images as they are overparameterized. Fine-tuned ResNet-50 model is used for the tasks of diabetic retinopathy detection, lung disease classification, malignant melanoma detection. Implemented attacks such as FGSM, BIM, PGD and CW reduce accuracy of the models close to 0 with very small perturbation value of 0.6/255. BIM and PGD attacks were successful in reducing accuracy and AUC to 0 against diabetic retinopathy detecting

model. For Dermoscopy model, accuracy reduces to <1 against the attacks. Results shows that FGSM is the least successful attack compared to other implemented attacks. They also argue that same complex biological structures present in medical images also help with detection of adversarial attacks. Methods such as Kernel Density (KD) and Local Intrinsic Dimensionality (LID) are used to detect adversarial attacks. Results of attack detection in medical images (99% AUC) is significantly better than detection in natural images (less than 80% AUC).

Li et al. [28] introduce a novel detection strategy for adversarial images. They show that the small noise in adversarial examples is hard to detect with naked eyes, but it is amplified after multiple convolution-pooling layers of a CNN. Based on this observation, they add a detection module between final fully connected layer and output layer of the CNN. Universal multivariate Gaussian model (MGM) is used for detection. They compare their approach with Isolation Forest (ISO) and One-class SVM (OCSVM) methods. Performance of the method is evaluated using DenseNet-121 trained on chest X-ray dataset, against attacks such as FGSM, Basic Iterative Method (BIM) and Projected Gradient Descent (PGD). Proposed approach yields significant improvement in F1 score and AUROC values. White box attacks are implemented directly on DenseNet-121 model while black attacks make use of surrogate ResNet-50 model to craft attacks and transfer it to original DenseNet-121 model.

2.3 Related works that defend against Adversarial Attacks in Medical Imaging Applications

Ren et al. [29] implement the network model derived from Cascaded Anisotropic Convolutional Neural Network. This network is then appended with task reorganization

module and adversarial defense module. Adversarial samples are generated using FGSM attack. These adversarial samples are then used for adversarial training. Final model improves dice score to 83.12% compared to existing dice score of 81.95% for U-Net and 81.88% for V-Net.

Calivá et al. [30] perform adversarial attacks on reconstruction network that generates MRIs. This work uses Fast-MRI dataset consisting of training/validation split of 973/199 volumes of knee MRI. They make use of fast negative features to perform adversarial attacks. Fast negative features are small perceptible features which are present in original images but disappear from the reconstructed images. This suggests that the network suffers from hallucination. False Negative Adversarial Features (FNAF) attack is used to find small features which are difficult for the network to reconstruct. This small abnormalities are then used in the adversarial training to make network robust and reduce its sensitivity to adversarial features.

2.4 Limitations of Existing Works

Research works discussed in above section show that adversarial attacks present serious threat to medical image processing system. However, many of these works have obvious limitations. Ma et al. [18] , Finlayson et al. [27] and other use models that are pretrained on ImageNet dataset which is a natural images dataset. Fine tuning a model on medical image dataset does not represent the real world medical imaging system. For better understanding of the effect of adversarial attacks, neural networks entirely trained on medical images are needed. This issue is addressed in this research work, all of the models used are trained on 3D brain MRI from scratch.

Some research works ([18], [24]) only implement white box attacks which assumes that attacker has some knowledge of the system. Adversarial attacks presented as black box attacks ([27] , [28]) are implemented by training a surrogate models making use of transferability property of the attacks. While possible in real world, it does not entirely cover all the possible black box attack threats. Black box attacks without surrogate models are still possible where attacker makes use of access to existing model's prediction. We cover this scenario with Resource Efficient Decision-based (RED) attack [39].

Research works that implement defense against adversarial attacks ([29], [30]) make use of adversarial training as a defense. While promising, adversarial training has a strong coupling with the adversarial attack algorithm used to generate adversarial examples. This results in a model that is only robust against adversarial attacks used during training. Hence, this work proposes a defense solution that is independent of the adversarial attack algorithms.

CHAPTER III

RESEARCH OBJECTIVE

The main objective of this work is to demonstrate the effect of adversarial attacks on MRI processing systems and to propose a novel approach that delivers an MRI processing system that is more robust than existing state-of-the-art works. The effect of adversarial attacks on prediction of MRI processing systems is demonstrated by implementing white box attacks such as FGSM and l_0 attack [40] as well as black box attacks such as RED attack [39]. We wanted to demonstrate that by generating perturbed inputs with very small amount of noise, prediction of these medical image processing systems can change from highly accurate to completely unreliable.

To design a robust MRI processing system that protects against adversarial attacks, we combine SRGAN and CNN model. Our hypothesis is that SRGAN can be used to remove adversarial perturbation from the input. Making use of SRGAN's ability of learning high frequency features and CNN's ability of high accuracy predictions, we aim to remove small perturbation added to the original input by adversarial attacks and recover an MRI that is closest to the original input generated by SRGAN.

CHAPTER IV

METHODOLOGY

Figure 1 shows step by step process of this research work. Preprocessing consists of brain MRI segmentation using Spatially Localized Atlas Network Tiles (SLANT) [41], which has two outputs, normalized brain MRI and anatomical features. Output of preprocessing is used to train CNN and Context Aware model proposed in [40] and the new proposed model which uses SRGAN [42] and CNN. After training, Gradient-based l_0 and l_∞ adversarial attacks are performed on all three models and compare the deviation caused by these attacks. Hypothesis for our novel approach is that SRGAN can be used to remove adversarial noise from the perturbed input.

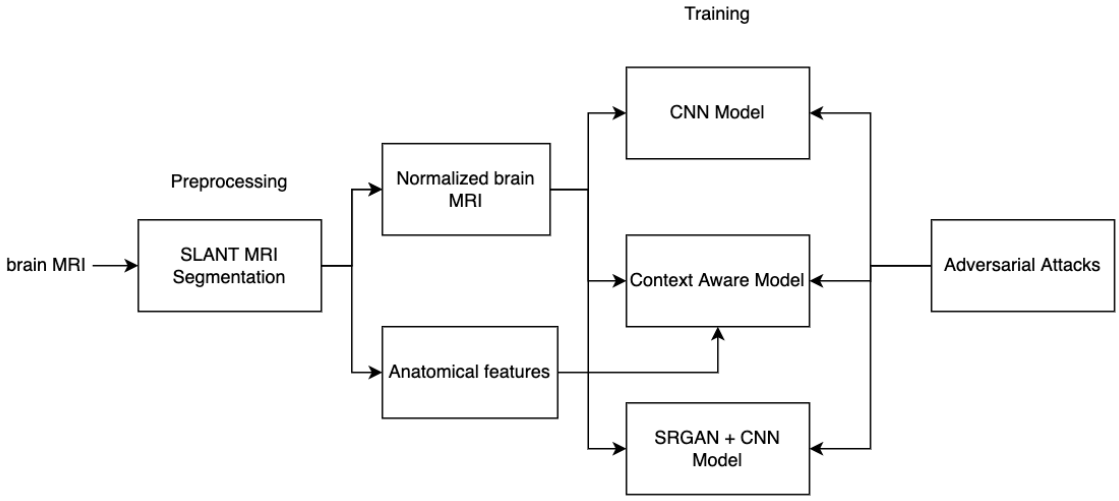

Figure 2 Methodology of the experiment

4.1 Dataset

The MRI dataset has a total of 2395 T1-weighted brain MRIs from three datasets: Autism Brain Imaging Data Exchange I (ABIDE I) [43], Attention Deficit Hyperactivity Disorder (ADHD) 200 Sample [44], FCON1000 from '1000 Functional Connectomes' Project [45]. These 2395 MRIs consist of 1102, 171, and 1122 MRIs from ABIDE I, ADHD 200 and FCON100 respectively. Subject age range vary from 7 to 85. After preprocessing with SLANT, each MRI has dimension of (172, 220, 156) and size of ~14MB. Out of these, patient MRIs with age range of 8 to 30 are selected as it helps obtaining more balanced data.

4.2 Preprocessing: Brain MRI Segmentation

MRI preprocessing aims to normalize each MRI in the dataset as well as obtain anatomical features from the image. Normalization is necessary as each patient's brain size is different and MRIs have different resolution and orientation depending on the MRI scanner used. Anatomical features are used as an input to Context aware model. To obtain anatomical features, MRI segmentation is performed. There are various methods to perform segmentation on brain MRIs. They can be categorized in 1) Gray level features-based methods, 2) Texture features-based methods, 3) Model-based segmentation and 4) Atlas-based segmentation. Atlas based segmentation methods can be further divided into Single Atlas Segmentation and Multi Atlas Segmentation (MAS). Out of these, atlas-based method is the latest and provides better accuracy. Figure 3 shows a typical multi atlas segmentation pipeline. Typical flow of MAS pipeline consists of acquiring multiple atlases segmented by human experts, registering target image to these atlas and solve label conflicts resulting from label estimation via label fusion.

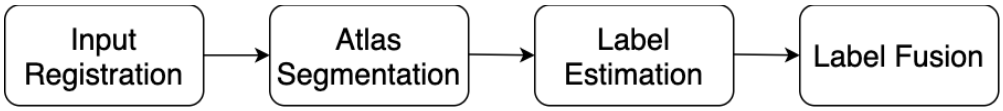

Figure 3 Multi Atlas Segmentation (MAS) Pipeline

There are many works involving multi atlas-based segmentation method, which use state-of-the-art CNNs to segment brain MRIs [46], [47]. In this work, SLANT [41] method is used for multi atlas-based segmentation. One of the limiting factor of traditional MAS methods is high computational cost. To reduce it SLANT limits the number of atlases used in registration. Deep learning algorithms implementing MAS use CNNs in place of label fusion, which eliminates the voting-based mechanism of fusion and uses learned features of CNNs instead.

SLANT divides input brain MRI input into 8 (non-overlapping) or 27 (overlapping) 3D tiles. Each tile is then trained on a separate CNN. This approach simplifies the tasks of each network and helps the network learn features of the small tile better compared to a single network. Since SLANT is dividing input into smaller parts, input image and its features need to be consistent, for this SLANT uses affine registration as a first step. Using Montreal Neurological Institute (MNI-305) template for affine registration, each brain MRI is transformed to dimension of (172, 220, 156). Atlases used in SLANT are manually labeled using brainCOLOR protocol, hence SLANT segments entire brain MRI into same 133 labels used in the protocol. Figure 4 shows preprocessing pipeline using SLANT. For preprocessing, each 3D brain MRI is fed to slant singularity container (https://github.com/MASILab/SLANTbrainSeg). SLANT first applies affine normalization to brain MRI. This normalized MRI is used as an input to all three models used in this experiment. SLANT outputs anatomical features which are being used as an input to Context aware model in addition to normalized MRI.

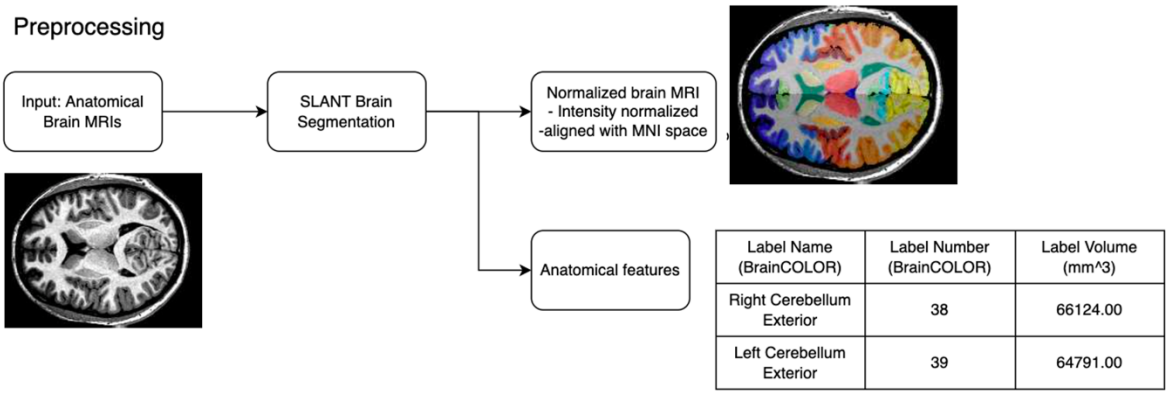

Figure 4 Preprocessing pipeline using SLANT.

4.3 Base Network Models

This work aims to improve the robustness of medical imaging diagnostic process which uses DL models. However, before discussing the proposed approach, it is pertinent to understand the architecture of the DL models such as CNN, Context-aware model [40] that are used as a benchmark and SRGAN [42] model which is used as a submodule for the proposed approach.

4.3.1 Convolutional Neural Network (CNN)

CNN considered in this work has a standard structure shown in Figure 5. First module of network that processes visual features is made up of five layers of convolutional layers with increasing filters size that range from 8 to 128. Each convolutional layer is followed by max-pooling layer. After convolutional stage, there are two dense layers and one activation layer. ReLU is the activation function for hidden dense layers.

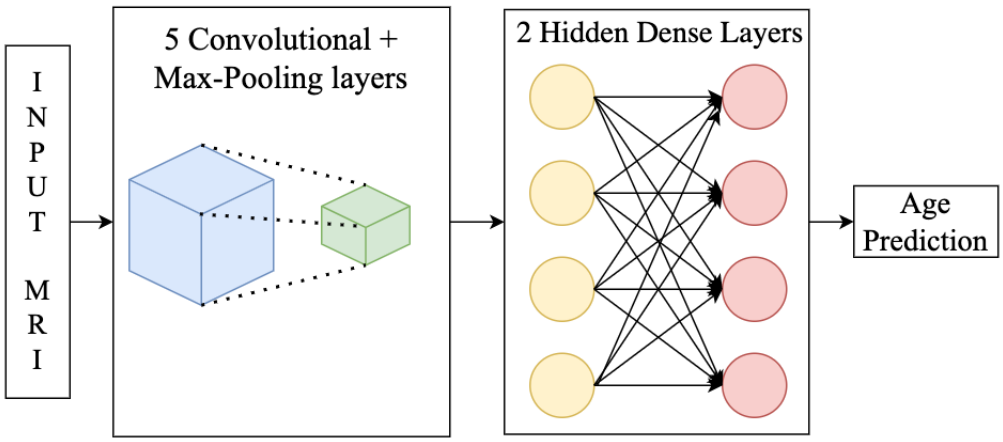

Figure 5 CNN Architecture

4.3.2 Context Aware Model

Context Aware model is a hybrid model which extends CNN shown in Figure 5. Context aware model has two branches, one to processes convolutional features which includes convolution layers from CNN and the other branch that processes anatomical features obtained by SLANT brain MRI segmentation discussed in section 4.2. Convolutional branch processes MRI features with same five convolutional layers as CNN and appends two sets of dropout layer, dense layer, and activation layer. Anatomical features branch only has activation layer. Both branches are concatenated using concatenate layer followed by one dropout layer and two more sets of dense and activation layer. Input of anatomical branch is normalized using min max scaler and standard scaler into value of -1 to 1. To match this magnitude, output of convolutional branch is also normalized to value of -1 to 1. Context Aware model architecture is shown in Figure 6.

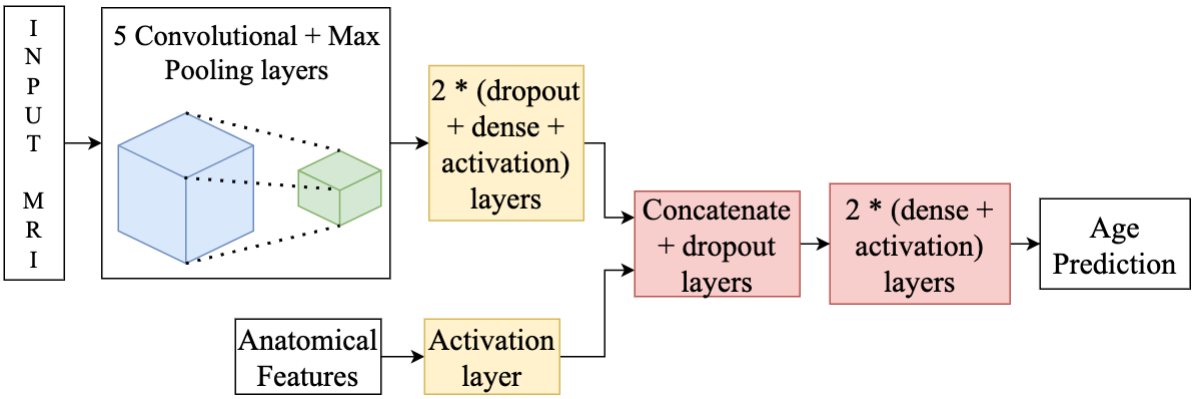

Figure 6 Context Aware Model Architecture

4.3.3 Super-Resolution Generative Adversarial Network (SRGAN)

Problem of recovering high-frequency details from a low resolution image has been solved by SRGAN [48] and it is designed with residual blocks. It achieves photo realistic super resolution images. Residual blocks allow for deeper neural network that don't suffer from the problem of exploding and vanishing gradients. For this research work, SRGAN proposed for brain MRIs [42] is used. The Generator includes 6 residual blocks. Each residual block is made up of convolutional layer, batch norm layer and ReLUs. Residual layers are followed by upsampling blocks and final convolution layer, that generates the high resolution image. Discriminator is made up of convolutional layers and dense layers. Figure 7 illustrates SRGAN architecture comprised of Generator and Discriminator.

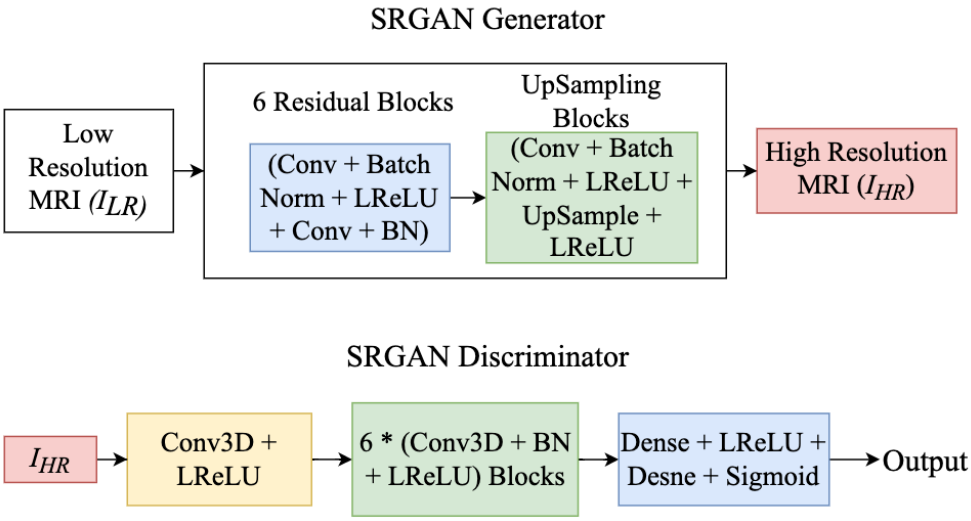

Figure 7 SRGAN Architecture

4.4 White Box Attacks

Attacks implemented in this experiment are gradient-based attacks based on fast gradient sign method (FGSM) attack proposed in [14], [49] and its modification presented in [40].

4.4.1 FGSM/l_∞ Attack

FGSM attack exploits linear behavior of neural networks in high-dimensional spaces to create the adversarial image. Intuition behind FGSM attack can be explained by a digital image that uses 8 bits per pixels. This image will discard any information that can be represented only using more than 8 bits. Similarly, FGSM attack aims to change the image by a magnitude that is big enough for neural network to change prediction but not enough for human eyes. Consider neural network as a function F, which take can input x and maps it to output y. The sign of the gradient of $F(x)$ can be used to create the adversarial input x' as,

$$x' = x + \epsilon * sign(\nabla(F(x))) \quad (3)$$

where $\nabla(F(x))$ is gradient of $F(x)$ with respect to x and ϵ is l_∞ distance or noise multiplier.

Algorithm to implement the l_∞ attack is illustrated in Algorithm 1.

Algorithm 1 FGSM/l_∞ attack

input: model F, array of l_∞ distances E, legitimate image x
output: adversarial input x'
$y = F(x)$
for ϵ **in** E **do**
    $direction = sign(\nabla(F(x), y))$
    $pertubation = clip_{(min,\max)}(\epsilon * direction)$
    $x' = x + perturbation$
end for

4.4.2 l_0 Attack

The l_0 attack aims to achieve maximum deviation in prediction by changing the minimum number of pixels. For each iteration, the pixel in the input image with maximum gradient value is changed. This perturbed image is checked for its prediction. If the changed pixel causes deviation, that change is kept otherwise that pixel is restored to its original value. To determine, which pixel to change is determined using the gradient of input image with respect to output. Pixel with highest gradient value is changed to its maximum value and the tested for prediction. This makes l_0 attack highly efficient causing high deviation with fewer than 500 iterations. Algorithm to implement the l_0 attack is shown in Algorithm 2.

Algorithm 2 l_0 attack

input: model F, legitimate image x, upper bound of l_0 distance ϵ, range of possible pixel values $V = [v_1, v_2, \dots, v_n]$
output: adversarial input x'
$x' = x$
$gradient = \nabla(F(x), y)$
while $i < \epsilon$ **do**
    $pos = argmax(|gradient|)$
    for k **in** V **do**
        $x'[pos] = k$

if not $F(x') > F(x)$ **then**
$x'[pos] = x[pos]$
end if
end for
end while

4.5 Black Box Attacks

In most real life scenarios, the attacker does not have full knowledge of the model and its parameters. This makes black box attacks more realistic. Black box attacks treat model as the or- acle and only use predictions from the model to attack the system. In this work, robustness of all neural network models is tested against state-of-the-art black box attack - Resource Efficient Decision-based (RED) Attack [39].

4.5.1 Resource Efficient Decision-based (RED) Attack

RED attack is a black box attack that is more efficient than existing black box attacks such as DeepFool [36]. In the original paper, RED attack was performed on CNN models trained on the CIFAR-10 and the GTSR datasets. This work aims to perform the same attack on the CNN, Context aware and SRGAN + CNN model trained on brain MRIs. This implementation of the attack aims to change prediction of all test dataset MRIs to maximum prediction value (age ~ 30) irrespective of their original value.

RED attack relies on three hyperparameters:

$\boldsymbol{d_{min}}$**:** represents maximum allowed perturbation for each pixel while finding boundary estimated image. Typical value of d_{min} for an RGB image (pixel value ranges from 0 to 255) is anywhere between 1 to 5. Normalized brain MRI pixels have magnitude range of approximately -4 to 17. Considering this, value of d_{min} is changed to 0.5 in this experiment.

n: defines total number of pixels that are perturbed while generating new adversarial MRI ($ii2MRI$) from boundary estimated MRI. Range of n for RGB image is from 5 to 50 for the image of 900 pixels. MRI is a 3D structure and has close to 2.7 million pixels. Hence value of n here is scaled to 10000.

theta (θ): It is multiplied with maximum value of any pixel can have and represents the amount of noise that will be added to adversarial image during gradient estimation stage. Typical value of θ ranges from 0.0196 to 0.196. In current implementation, value of θ is 0.196.

Each iteration of RED attack is performed using three algorithms: (i) Boundary Estimation, (ii) Gradient Estimation and (iii) Efficient Update. In the first step, boundary estimation (Algorithm 3) aims to perturb source image, such that it lies on the boundary of the source and target class prediction space distribution.

Algorithm 3 RED Attack – Boundary Estimation

input: age classifier model *F*, MRI from source class $sourceMRI$, MRI from target class $targetMRI$, maximum allowed perturbation d_{min}
output: Adversarial MRI $iiMRI$

$iiMRI = (sourceMRI + targetMRI) / 2$
$k = F(iiMRI)$
$delta = max(sourceMRI - iiMRI)$

while $(delta > d_{min})$ **do**
    if $F(targetMRI) \neq k$ **then**
        $sourceMRI = iiMRI$
    else
        $targetMRI = iiMRI$
    end if
    $iiMRI = (sourceMRI + targetMRI) / 2$
    $k = F(iiMRI)$
    $delta = \max(sourceMRI - iiMRI)$
end while

Gradient estimation (Algorithm 4) takes boundary estimated MRI as an input and changes n random pixels of the image to the maximum pixel value. After introducing random perturbation, the direction in which MRI has moved from boundary is checked. If direction is closer to source MRI, the direction of the change is reversed and vice versa.

Algorithm 4 RED Attack – Gradient Estimation

input: MRI from source class $sourceMRI$, MRI from target class $targetMRI$, boundary estimated MRI $iiMRI$, number of pixels to perturb n, perturbation multiplication factor $theta$
output: perturbed MRI $ii2MRI$, gradient direction g

$I_0 = zeros$
Set n random pixels of I_0 to their maximum value
$ii2MRI = iiMRI + (theta * I_0)$
$diff1 = iiMRI - sourceMRI$
$diff2 = ii2MRI - sourceMRI$

if $(diff2 > diff1)$ **then**
    $g = 1$
else if $(diff2 < diff1)$ **then**
    $g = -1$
else
    $g = 0$
end if

Efficient update generates new adversarial MRI, *inewMRI* based on gradient direction g and jump size j. This newly generated MRI will be optimized for its l_2 distance from source MRI. Total number of update iterations are constrained by *maxCount*, after which boundary estimated MRI is considered as the adversarial MRI.

Algorithm 5 RED Attack – Efficient Update

input: MRI from source class $sourceMRI$, MRI from target class $targetMRI$, boundary estimated MRI $iiMRI$, randomly perturbed image $ii2MRI$, gradient direction g, jump size j, maximum iterations max$Count$
output: perturbed MRI $iNewMRI$

$delta = g * (ii2MRI - iiMRI)$
$iNewMRI = iiMRI + (j * delta)$

$d1 = \sum (iNewMRI - sourceMRI)^2$
$d2 = \sum (iiMRI - sourceMRI)^2$
$count = 0$
while $d1 > d2$ **and** $count < maxCount$ **do**
    $iNewMRI = iiMRI + (j * delta)$
    $d1 = \sum (iNewMRI - sourceMRI)^2$
    $count = count + 1$
    if $d1 > d2$ **then**
        $iNewMRI = iiMRI$
    end if
end while

Algorithm 6 describes how each step of the RED attack is used in a single iteration. First, we find an MRI that lies on boundary of source and target class distribution using boundary estimation described in Algorithm 3. Next, gradient estimation is done between boundary estimated MRI and randomly perturbed MRI (Algorithm 4). Efficient update step uses both MRI and returns, whichever is closest to source MRI. l_2 norm of output MRI is compared with minimum norm MRI found till that iteration. If l_2 has not decreased, we discard output of that iteration. The end result of the attack is an MRI that has minimum l_2 distance to MRI from the source class but CNN predicts it to target class.

Algorithm 6 RED Attack – Iteration

input: MRI from source class $sourceMRI$, MRI from target class $targetMRI$, maximum iterations *maxIterations*, age classifier model *F*
output: perturbed MRI $minNormMRI$

$beMRI$ = boundary_estimation($sourceMRI$, $targetMRI$) // Algorithm 3
$targetPrediction = F(targetMRI)$
$minNormMRI = beMRI$
$minl2Norm = \sum (minNormMRI - sourceMRI)^2$
while *i* < *maxIterations* **do**
    gd, $geMRI$ = gradient_estimation($sourceMRI$, $beMRI$) // Algorithm 4
    $euMRI$ = efficient_update($sourceMRI$, $beMRI$, $geMRI$, gd) // Algorithm 5
    $finalPrediction = F(euMRI)$
    $eul2Norm = \sum (euMRI - sourceMRI)^2$

if $eul2Norm < minl2Norm$ **then**
if $finalPrediction \neq targetPrediction$ **then**
$euMRI = \text{boundary\_estimation}(sourceMRI, targetMRI)$
end if
$minNormMRI = euMRIs$
$minl2Norm = \sum (minNormMRI - sourceMRI)^2$
end if
$i = i + 1$
end while

4.6 Proposed Hybrid Architecture: SRGAN with CNN

SRGAN excels at learning the mapping between low resolution image and high resolution image. This property of SRGAN is leveraged in proposed system of combined SRGAN and CNN to protect medical imaging system. SRGAN proposed in [42] is modified to work with brain MRI resolution of (172, 220, 156). Generator network in SRGAN is trained to generate high resolution image I_{HR} from low resolution image I_{LR}. In the training phase generator learns the features of training dataset which only contains legitimate input MRIs that are compressed. When we feed adversarial input to the combination of SRGAN and CNN, first adversarial input is compressed, which remove certain degree of noise. Since SRGAN generator is only trained on legitimate input, it generates closest legitimate input from the adversarial input and removes large degree of adversarial perturbation. As output of SRGAN is closest to legitimate input, CNN prediction will also be very close to the legitimate input. Below are the details of implementation:

4.6.1 Model Design

To train SRGAN generator, normalized MRIs from SLANT output are used. Low resolution images I_{LR} are obtained by scaling MRIs by a factor of 1/2. Gaussian noise is added to I_{LR} in order to improve generator performance. Sub-pixel convolution

configuration of the generator network is used for upsampling. Each MRI is divided into 8 patches while training to improve the efficiency and make training possible on a single GPU. Generator output is a high resolution image I_{HR} which has same resolution as the original image. Discriminator network is discarded after training. After training, entire dataset is evaluated on the generator network to obtain a new dataset made up of high resolution images. CNN model is trained using this dataset of high resolution images. Figure 8 shows flow diagram of our training process.

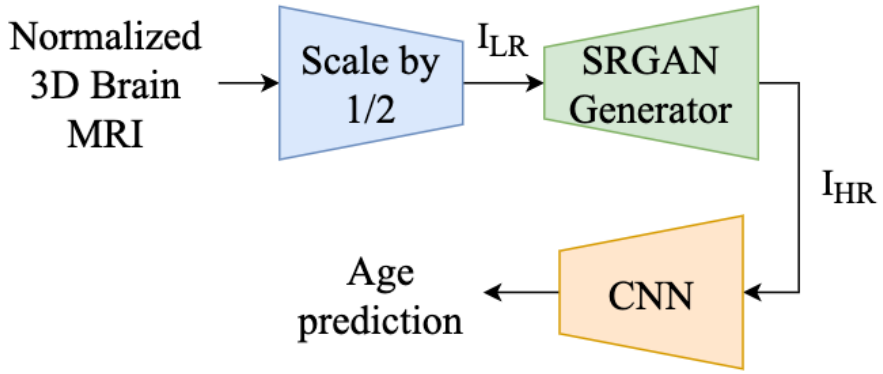

Figure 8 Proposed Architecture: Medical imaging system using SRGAN and CNN

4.6.2 Adversarial Attack Design for SRGAN+CNN Model

As proposed system is using two separate modules SRGAN and CNN, adversarial attack on the system needs to consider both modules while creating adversarial input. All three attacks implemented here considers both modules in the implementation. White box attacks FGSM and l_0 are implemented in two stages. In the first stage, adversarial input is created for SRGAN generator module. This adversarial input is saved as perturbed high resolution MRI I_{PHR} along with the gradient of generator module $\nabla_{SRGAN}(F(I_{LR}), I_{PHR})$. Second stage of the attack uses I_{PHR} (output of the first stage) to create adversarial input for CNN. FGSM and l_0 attack use gradient for attack direction as below,

$$direction = sign(\nabla_{SRGAN}(F(I_{LR}), I_{PHR}) * \nabla_{CNN}(F(I_{PHR}), \mathrm{y}) \quad (4)$$

Rest of the attack steps are similar to Algorithm 1 and Algorithm 2. RED attack follows similar approach where each MRI prediction is passed through SRGAN as well as CNN module.

CHAPTER V

EXPERIMENTS AND RESULT

This experiment was performed on MRI dataset with the age group of 8 to 30. Data distribution for age group 8 to 30 is shown in Figure 9 (a). It suggests that data distribution is skewed. Random oversampling was used to balance the dataset. Balanced dataset is displayed in Figure 9 (b).

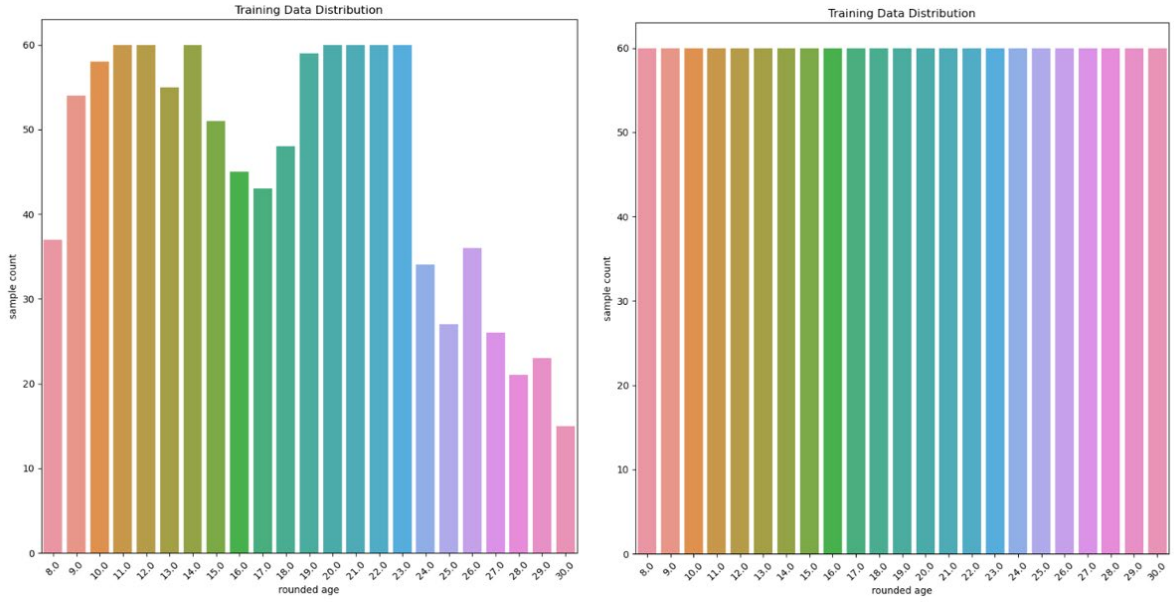

Figure 9 Training dataset distribution. a) unbalanced dataset b) balanced dataset

5.1 Model Training Results

All three models were trained on same training dataset for accurate comparison. As ML algorithms are computation heavy algorithms, all models in this experiment were trained on Ohio Supercomputer (OSC) [50]. CNN is trained with learning rate of 0.001 and optimizer selected is Adam. Due to the large size of a single brain MRI, CNN is trained parallelly on two NVIDIA V100S 32GB using TensorFlow 2's distributed MirroredStrategy. Large file size limits batch size to maximum of 24. Training took around 12 hours for 65 epochs using training dataset that consists of 1255 MRIs. Context aware model is trained on pretrained CNN and anatomical features. This makes training time for Context aware model much shorter, which is 30 minutes for 65 epochs.

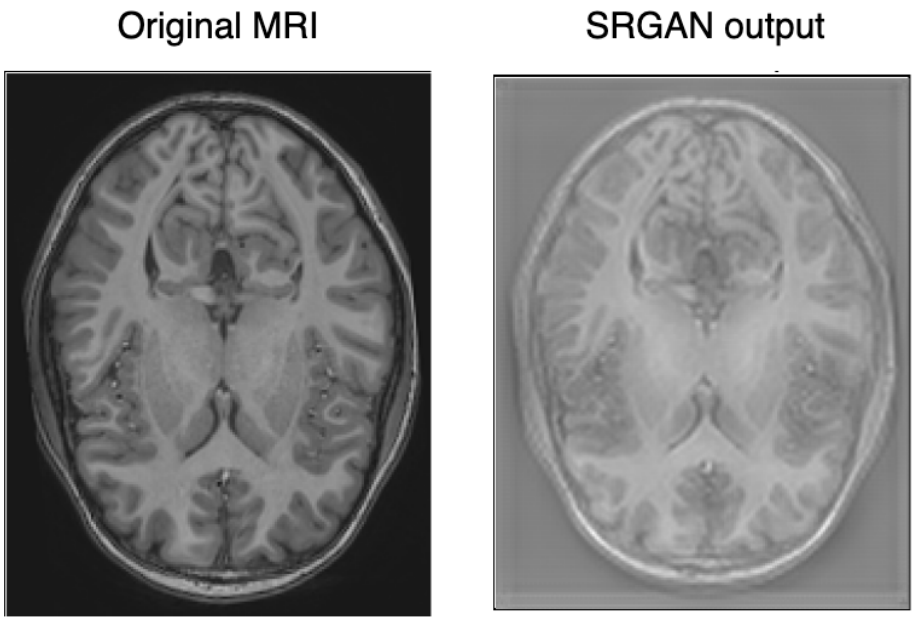

Figure 10 Reconstruction of brain MRI using SRGAN

For SRGAN implementation, SRGAN architecture presented in [42] was modified to fit normalized MRI resolution of the dataset. SRGAN is trained with batch size of 1 MRI, which is divided into 8 patches. So, each epoch trains on 8 patches of a single MRI.

This helps in reducing the computation and memory requirement. As a result, SRGAN required only a single 32 GB NVIDIA V100s GPU. Training took approximately 8 hours for dataset of 275 MRIs. Input brain MRI is downscaled by the factor of 2 and SRGAN increases its resolution by 2 providing original image back. Average peak signal to noise ratio (PSNR) is 34.25 and PSNR variance is 3.56. SRGAN output is shown in Figure 10. After SRGAN training is finished, entire dataset is evaluated on SRGAN. The output of SRGAN is then used to train CNN. CNN training and configuration are kept similar to CNN discussed above.

Performance of each model is measured in terms of Root Mean Square Error (RMSE) which is displayed in Table 2. Results show that prediction RMSE of hybrid SRGAN and CNN model is on a par with existing CNN and Context aware models.

Model	Prediction RMSE (Training)
CNN	3.83
Context Aware	3.71
SRGAN + CNN	3.91

Table 2 Training RMSE values

5.2 White Box Attacks (l_∞ and l_0 attacks)

Robustness of the model against adversarial attacks is measured in terms of deviation ($|F(x') - F(x)|$). Noise levels are selected in such a way that they are imperceptible to human eyes. Similar to [40], FGSM/l_∞ attack was performed with six l_∞ noise level: 0.0001, 0.0002, 0.0005, 0.001, 0.002, and 0.005. As per definition of l_∞ norm, FGSM attack limits the maximum amount of noise added to the input MRI. l_0 attack limits the number of pixels changed from their original value. Both attacks try to perturb input in such a way that predicted age is maximized.

Perturbation value (multiplication factor ϵ)	Prediction Error in various age groups (in years)								
	CNN model			Context Aware model			SRGAN + CNN model		
	<15	15-22	>22	<15	15-22	>22	<15	15-22	>22
0.0001	24.79	25.99	24.32	11.57	14.78	15.41	0.08	0.09	0.09
0.0002	40.16	42.35	39.28	19.70	24.62	25.45	0.16	0.19	0.19
0.0005	64.62	67.92	63.02	33.58	41.78	42.88	0.39	0.47	0.48
0.001	82.72	87.14	81.29	42.72	54.15	55.55	0.78	0.93	0.95
0.002	98.05	104.18	97.62	46.84	60.30	62.57	1.53	1.85	1.90
0.005	112.59	121.01	114.11	44.66	58.86	59.58	3.67	4.53	4.60

Table 3 Deviation in prediction caused by FGSM attack

Table 3 shows prediction deviation in various age groups caused by the FGSM attack in all network models. Results show that even the lowest noise value of 0.0001 causes deviation of close to 25 years across all age groups in CNN. This level of error would make the system unreliable. With increased noise levels, performance of the system only gets worse reaching prediction error of more than 110 years across all age groups for noise level 0.005. Context aware model is more robust than CNN thanks to its use of anatomical features. For most age groups, prediction error in context aware model reduces by close to half, resulting in overall 44.6% less deviation across entire dataset. This is a significant improvement over CNN model, but average deviation is still at 13.92 for noise level of 0.0001 and 54.37 for noise level 0.005. System would still be unusable at this level. Proposed approach of hybrid SRGAN and CNN models improves considerably for CNN and Context aware model. Average deviation is 0.09 for noise level of 0.0001 and 4.27 for noise level of 0.005. This shows that SRGAN+CNN model is 99.68% and 99.42% more robust than CNN and Context aware model respectively for noise level of 0.0001. Similarly, for the noise level of 0.005, it is 96.74% and 93.24% more robust than CNN and Context aware model respectively.

l_0 attack was implemented with range of l_0 noise from 5 pixels to 25 pixels in increments of 5. Each MRI in the dataset has dimension of (172, 220, 156), which is 5.9 $*$

10^6 pixels. Table 4 shows results of l_0 attack on all three models. This shows effectiveness of the l_0 attack, by only changing 5 pixels, we were able to add 2.51 years to the predicted age in CNN. Context aware is more robust against l_0 attacks compared to CNN but only by 25% compared to 43.74% against FGSM. SRGAN+CNN renders l_0 attack nearly ineffective for given noise values as show in Table 4. SRGAN+CNN model is 96.20% and 94.91% more robust than CNN and Context aware model respectively.

Table 3 and Table 4 show that prediction error increases with rising age. For FGSM attack, deviation increases with the age. Deviation pattern for l_0 attack is random but is generally higher in ranges of 15 to 22 and >22.

Perturbation value (number of pixels)	**Prediction Error in various age groups (in years)**								
	CNN model			**Context Aware model**			**SRGAN + CNN model**		
	<15	**15-22**	**>22**	**<15**	**15-22**	**>22**	**<15**	**15-22**	**>22**
5	2.65	2.77	2.12	1.14	2.08	2.42	0.16	0.12	0.12
10	4.48	4.74	3.72	2.15	3.26	4.26	0.20	0.16	0.15
15	6.12	6.55	6.54	3.28	4.66	5.61	0.25	0.21	0.19
20	7.66	8.01	6.54	4.33	5.82	6.56	0.27	0.24	0.21
25	9.03	9.47	7.76	5.31	7.51	7.78	0.30	0.27	0.25

Table 4 Deviation in prediction caused by l_0 attack

Figure 11 shows average deviation caused by FGSM attack based on various noise levels in all models. It shows how SRGAN+CNN model significantly improves the robustness across all noise levels. For small noise levels such as 0.0001, 0.0002, and 0.0005, FGSM attack is rendered ineffective against SRGAN+CNN, causing average deviation of 0.23 years. Even for the largest noise level 0.005, deviation in SRGAN+CNN model is 96.3% less than that of CNN and 92.1% less than the context aware model. Maximum average deviation in the proposed SRGAN+CNN model is 4.22, which is very close to prediction RMSE. Similarly, SRGAN+CNN is overall 95.87% more robust against l_0 attack. Maximum average deviation of the proposed hybrid SRGAN+CNN model is

0.27 years compared to 8.89 years in CNN and 6.73 years in context aware model. Results for l_0 attack is shown in Figure 12 .

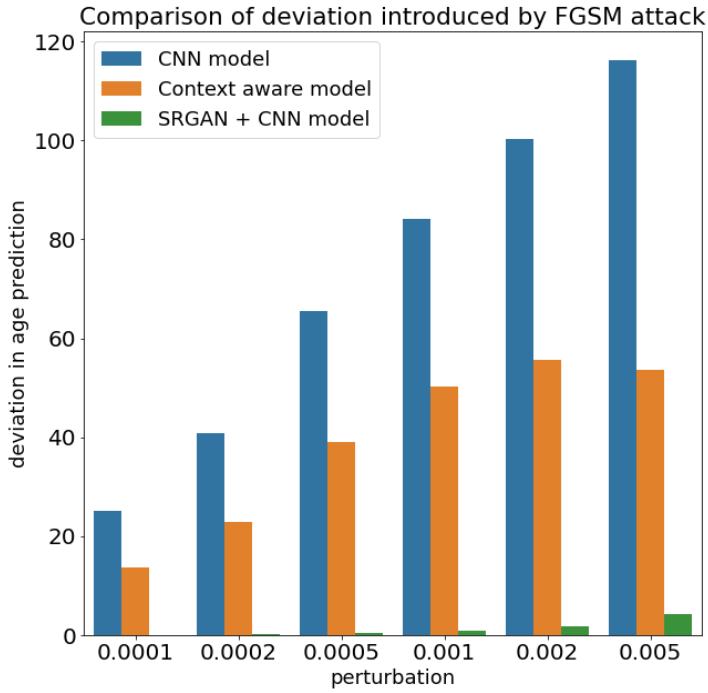

Figure 11 Comparison between models - FGSM attack

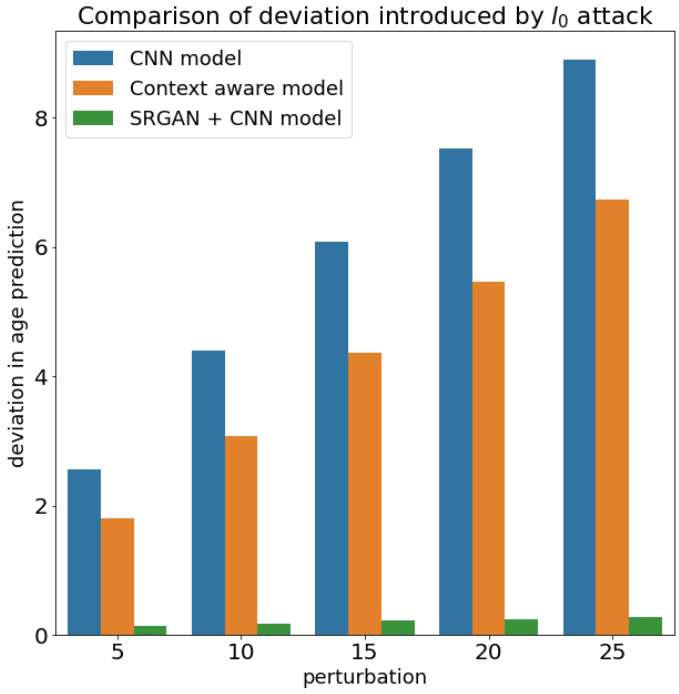

Figure 12 Comparison between models - l_0 attack

5.3 Black Box Attacks – RED Attack

The l_2 norm was used to measure the effectiveness of RED attack. These are the hyperparameter values used: $\delta_{min} = 0.5,\ n = 10000, \theta = 0.196$. Values of δ_{min} and θ is same as the related state-of-the-art work [39]. Value of n is scaled to match with close to 6 million pixels present in MRI compared to 900 pixel images of GTSR dataset used in the benchmark state-of-the-art work [39].

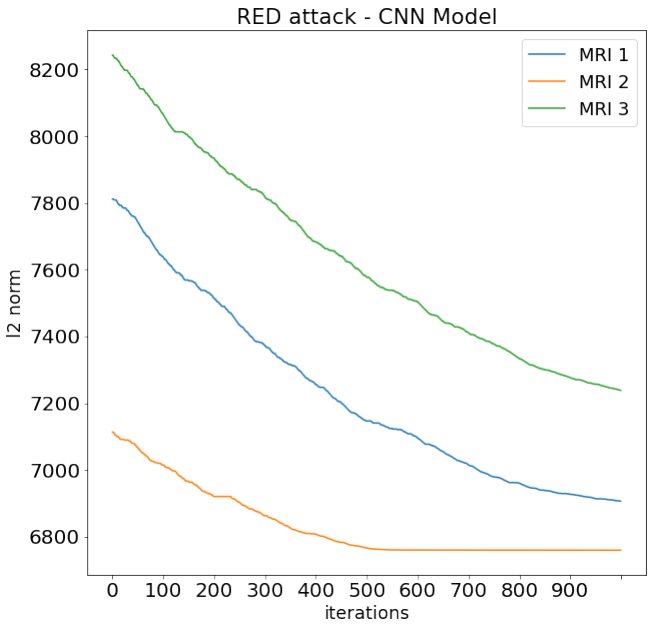

Figure 13 RED attack - CNN model

This implementation of RED attack aims to change the prediction of the source MRI to maximum age prediction present in the test dataset. RED attack was performed on CNN, Context Aware and SRGAN+CNN model for 1000 queries and results are presented in Figure 13**,** Figure 14 and Figure 15 RED attack - SRGAN+CNN model. Attack on Context aware model was performed without changing contextual features of the MRIs. For SRGAN+CNN model, each age prediction on MRI was done by first feeding MRI to SRGAN model and then feeding output of SRGAN model to CNN for age prediction.

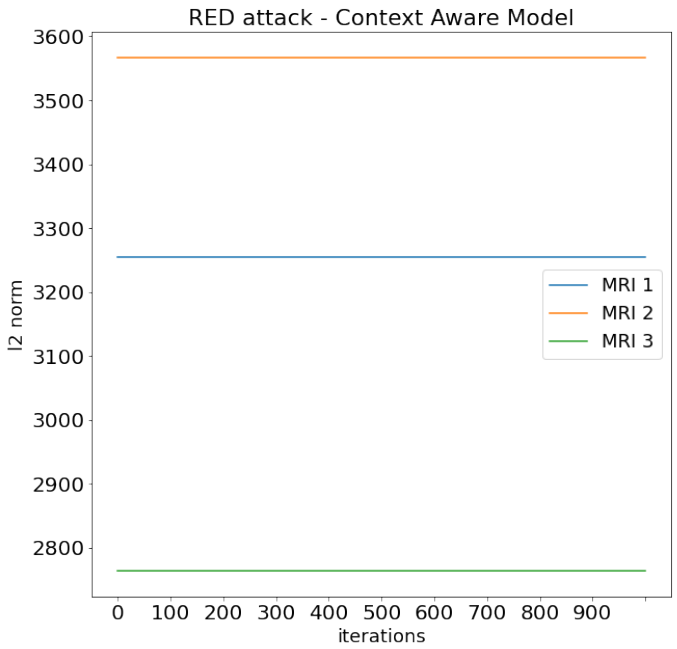

Figure 14 RED attack – Context Aware model

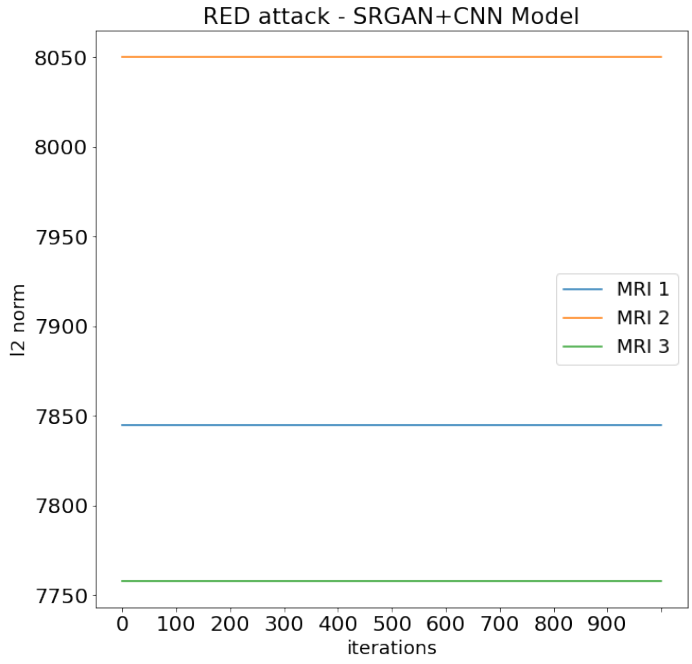

Figure 15 RED attack - SRGAN+CNN model

Effectiveness of the RED attack is measured by l_2 norm. It can be seen from the Figure 13(a) that as number of iterations increase, l_2 norm of perturbed image decreases

for CNN. This suggests that MRI is moving closer to source image. While for Context aware model, as shown in Figure 13(b) and SRGAN+CNN model, as show in Figure 14, l_2 norm stays constant during the attack. This suggests both models are equally robust where RED attack fails to find an image that lies on the source/target boundary but is closer to source MRI.

CHAPTER VI

LESSONS LEARNED

During the implantation and experiments, many technical issues occurred. This sections discusses some of the issues and how they were solved.

All the models and adversarial attacks in this work are implanted in Python using TensorFlow and Keras library. To train the network models, training and validation datasets needs to be divided into small chucks of respective batch size and then fed to neural network. Considering a single MRI is around 17 MB in size, dataset cannot fit in RAM during runtime. This meant that dataset has to be read from the disk for each batch. To fulfill this requirement, custom dataset generator was designed. During the training of CNN, batch size could not be increased more than 16 due to size of the MRI for a machine with single 32GB GPU. To address this issue, TensorFlow's MirroredStrategy from the distribute module was used. This allowed increase of batch size up to 24. Batch size value of 24 also helped solve the problem of exploding and vanishing gradient that neural networks suffer from.

Early implementations of the setup did not match the prediction RMSE of the related state-of-the-art work [40]. For this, entire setup had to be redesigned for hyperpar-

meter optimization. Current implementation includes a JSON configuration file under each experiment folder where hyperparameters such as learning rate, batch size, oversampling, variable learning rate can be changed. Age range of patient was also made narrower and random oversampling was applied to meet the desired RMSE.

CHAPTER VII

FUTURE WORK RECOMMENDATIONS

Further research in network architectures and model inputs can help make MRI processing systems more robust. SRGAN and CNN architectures can be improved using newly proposed architectures in natural image processing. SRGAN model can also be combined with context aware model to make system more robust. In current implementation, CNN training was limited at the batch size of 24 due to GPU size and large size of each 3D MRI image. CNN architecture can be improved by increasing batch size with the help of more capable GPUs.

SRGAN architecture [42] was also designed with GPU efficiency in mind, where each MRI was divided into 8 patches to training purposes. This can also be improved with more computation power and more GPU memory. Robustness of the system can be further tested on white box attacks such as Carlini-Wagner attack [15], JSMA attack [51] and black box attacks such as DeepFool [36].

One of the main vulnerability of neural networks is transferability property, where perturbed image for one model would also fool another model implementing similar functionality. Transferability property can also be tested on new SRGAN+CNN model by

training a surrogate model. While performing FGSM, l_0 and RED attack, anatomical features were kept constants for Context aware model due to high computational and time cost of obtaining the features for each MRIs. Obtaining new anatomical features after the MRI has been altered by the attack would provide more accurate robustness of the Context aware model.

CHAPTER VIII

CONCLUSION

This research aimed to test effectiveness of adversarial attacks against medical imaging systems using convolutional neural network and propose a new network architecture that protects against implemented attacks.

Results show that despite rapid progress in network architectures, neural networks are still very susceptible to small perturbations. Attacks such as FGSM and l_0 can change prediction of the models with perturbations that are imperceptible to human eyes. Black box attacks such as RED attack is very efficient at generating images that are closer to one class of image while keeping model prediction to entirely different class.

The proposed novel SRGAN-based approach to improve the classification robustness of a CNN against adversarial attacks is very effective. To simulate real word threat models, white box as well as black box attacks were performed against all three models that we implemented. Results suggest that SRGAN is indeed successful in removing small perturbations from the MRI and moving it close to original image. Combination of SRGAN and CNN improves system robustness by more than 95% against white box attacks compared to existing CNN and Context aware architectures. For black-

box attacks, SRGAN+CNN and Context aware model were equally more robust compared to regular CNN architecture.

REFRENCES

[1] J. B. Heaton and N. Polson, “Deep Learning for Finance: Deep Portfolios,” *SSRN Electronic Journal*, 2016, doi: 10.2139/ssrn.2838013.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical Text-Conditional Image Generation with CLIP Latents,” Apr. 2022.

[3] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” Jun. 2019.

[4] R. Valiente, M. Zaman, S. Ozer, and Y. P. Fallah, “Controlling Steering Angle for Cooperative Self-driving Vehicles utilizing CNN and LSTM-based Deep Networks,” in *2019 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, Jun. 2019, pp. 2423–2428. doi: 10.1109/IVS.2019.8814260.

[5] L. Chen, S. Wang, W. Fan, J. Sun, and S. Naoi, “Beyond human recognition: A CNN-based framework for handwritten character recognition,” in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, IEEE, Nov. 2015, pp. 695–699. doi: 10.1109/ACPR.2015.7486592.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” Dec. 2015.

[7] Andrej Karpathy, “What I learned from competing against a ConvNet on ImageNet,” Sep. 02, 2014.

[8] R. and C. R. and C. D. and J. K. and I. W. and D. T. Reynolds, “The Complexities of Physician Supply and Demand: Projections From 2019 to 2034,” Jun. 2021, doi: 10.13140/RG.2.2.29404.92808.

[9] A. Urushibara *et al.*, “Diagnosing uterine cervical cancer on a single T2-weighted image: Comparison between deep learning versus radiologists,” *Eur J Radiol*, vol. 135, p. 109471, Feb. 2021, doi: 10.1016/j.ejrad.2020.109471.

[10] M. T. Hagos and S. Kant, “Transfer Learning based Detection of Diabetic Retinopathy from Small Dataset,” May 2019.

[11] T. Nemoto *et al.*, “Simple low-cost approaches to semantic segmentation in radiation therapy planning for prostate cancer using deep learning with non-contrast planning CT images,” *Physica Medica*, vol. 78, pp. 93–100, Oct. 2020, doi: 10.1016/j.ejmp.2020.09.004.

[12] W. Gale, L. Oakden-Rayner, G. Carneiro, A. P. Bradley, and L. J. Palmer, “Detecting hip fractures with radiologist-level performance using deep neural networks,” Nov. 2017.

[13] “FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems,” Apr. 11, 2018. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye (accessed Sep. 04, 2022).

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” Dec. 2014.

[15] N. Carlini and D. Wagner, “Towards Evaluating the Robustness of Neural Networks,” Aug. 2016.

[16] C. Szegedy *et al.*, “Intriguing properties of neural networks,” Dec. 2013.

[17] T. Tanay and L. Griffin, “A Boundary Tilting Persepective on the Phenomenon of Adversarial Examples,” Aug. 2016.

[18] X. Ma *et al.*, “Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems,” Jul. 2019, doi: 10.1016/j.patcog.2020.107332.

[19] “NHE Fact Sheet.” https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet (accessed Sep. 04, 2022).

[20] G. R. Machado, E. Silva, and R. R. Goldschmidt, “Adversarial Machine Learning in Image Classification: A Survey Towards the Defender’s Perspective,” Sep. 2020, doi: 10.1145/3485133.

[21] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical Black-Box Attacks against Machine Learning,” Feb. 2016.

[22] J. Dvorak, J. George, A. Junge, and J. Hodler, “Age determination by magnetic resonance imaging of the wrist in adolescent male football players,” *Br J Sports Med*, vol. 41, no. 1, pp. 45–52, Oct. 2006, doi: 10.1136/bjsm.2006.031021.

[23] D. A. Wood *et al.*, “Accurate brain-age models for routine clinical MRI examinations,” *Neuroimage*, vol. 249, p. 118871, Apr. 2022, doi: 10.1016/j.neuroimage.2022.118871.

[24] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, “Generalizability vs. Robustness: Adversarial Examples for Medical Imaging,” Mar. 2018.

[25] G. Cheng and H. Ji, “Adversarial Perturbation on MRI Modalities in Brain Tumor Segmentation,” *IEEE Access*, vol. 8, pp. 206009–206015, 2020, doi: 10.1109/ACCESS.2020.3030235.

[26] G. Qi, L. Gong, Y. Song, K. Ma, and Y. Zheng, “Stabilized Medical Image Attacks,” Mar. 2021.

[27] S. G. Finlayson, H. W. Chung, I. S. Kohane, and A. L. Beam, “Adversarial Attacks Against Medical Deep Learning Systems,” Apr. 2018.

[28] X. Li and D. Zhu, “Robust Detection of Adversarial Attacks on Medical Images,” in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, IEEE, Apr. 2020, pp. 1154–1158. doi: 10.1109/ISBI45749.2020.9098628.

[29] X. Ren, L. Zhang, Q. Wang, and D. Shen, “Brain MR Image Segmentation in Small Dataset with Adversarial Defense and Task Reorganization,” Jun. 2019.

[30] F. Calivá, K. Cheng, R. Shah, and V. Pedoia, “Adversarial Robust Training of Deep Learning MRI Reconstruction Models,” Oct. 2020.

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” Feb. 2016.

[32] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 2017.

[33] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[34] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” May 2015.

[35] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, “The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation,” Nov. 2016.

[36] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “DeepFool: a simple and accurate method to fool deep neural networks,” Nov. 2015.

[37] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The Limitations of Deep Learning in Adversarial Settings,” Nov. 2015.

[38] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, “Adversarial Examples for Semantic Segmentation and Object Detection,” Mar. 2017.

[39] F. Khalid, H. Ali, M. A. Hanif, S. Rehman, R. Ahmed, and M. Shafique, “RED-Attack: Resource Efficient Decision based Attack for Machine Learning,” Jan. 2019.

[40] Y. Li, H. Zhang, C. Bermudez, Y. Chen, B. A. Landman, and Y. Vorobeychik, “Anatomical context protects deep learning from adversarial perturbations in medical imaging,” *Neurocomputing*, vol. 379, pp. 370–378, Feb. 2020, doi: 10.1016/j.neucom.2019.10.085.

[41] Y. Huo *et al.*, “3D whole brain segmentation using spatially localized atlas network tiles,” *Neuroimage*, vol. 194, pp. 105–119, Jul. 2019, doi: 10.1016/j.neuroimage.2019.03.041.

[42] I. Sanchez and V. Vilaplana, “Brain MRI super-resolution using 3D generative adversarial networks,” Dec. 2018.

[43] “Autism Brain Imaging Data Exchange I.” http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html (accessed Sep. 11, 2022).

[44] “The ADHD-200 Sample.” http://fcon_1000.projects.nitrc.org/indi/adhd200/ (accessed Sep. 11, 2022).

[45] “FCP Classic Data Sharing Samples”, Accessed: Sep. 11, 2022. [Online]. Available: http://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html

[46] A. V. Dalca, E. Yu, P. Golland, B. Fischl, M. R. Sabuncu, and J. Eugenio Iglesias, “Unsupervised Deep Learning for Bayesian Brain MRI Segmentation,” 2019, pp. 356–365. doi: 10.1007/978-3-030-32248-9_40.

[47] Z. Cui, J. Yang, and Y. Qiao, “Brain MRI segmentation with patch-based CNN approach,” in *2016 35th Chinese Control Conference (CCC)*, IEEE, Jul. 2016, pp. 7026–7031. doi: 10.1109/ChiCC.2016.7554465.

[48] C. Ledig *et al.*, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,” Sep. 2016.

[49] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” Jul. 2016.

[50] Ohio Supercomputer Center, “Ohio Supercomputer Center,” 1987. http://osc.edu/ark:/19495/f5s1ph73 (accessed Jun. 18, 2023).

[51] T. Combey, A. Loison, M. Faucher, and H. Hajri, “Probabilistic Jacobian-based Saliency Maps Attacks,” Jul. 2020.