Towards More Robust and Scalable Deep Learning Systems for Medical Image Analysis

Akshaj Yenumala
School of Computer Science
Georgia Institute of Technology
Atlanta, United States
akshaj@gatech.edu

Xinyue Zhang
Department of Computer Science
Kennesaw State University
Kennesaw, United States
xzhang48@kennesaw.edu

Dan Lo
Department of Computer Science
Kennesaw State University
Kennesaw, United States
dlo2@kennesaw.edu

***Abstract*—Deep learning (DL) has attracted interest in healthcare for disease diagnosis systems in medical imaging analysis (MedIA) and is especially applicable in Big Data environments like federated learning (FL) and edge computing. However, there is little research into mitigating the vulnerabilities and robustness of such systems against adversarial attacks, which can force DL models to misclassify, leading to concerns about diagnosis accuracy. This paper aims to evaluate the robustness and scalability of DL models for MedIA applications against adversarial attacks while ensuring their applicability in FL settings with Big Data. We fine-tune three state-of-the-art transfer learning models, DenseNet121, MobileNet-V2, and ResNet50, on several MedIA datasets of varying sizes and show that they are effective at disease diagnosis. We then apply the Fast Gradient Sign Method (FGSM) to attack the models and utilize adversarial training (AT) and knowledge distillation to defend them. We provide a performance comparison of the original transfer learning models and the defended models on the clean and perturbed data. The experimental results show that the defensive techniques can improve the robustness of the models to the FGSM attack and be scaled for Big Data as well as utilized for edge computing environments.**

***Index Terms*—Deep Learning, Adversarial Attack, Knowledge Distillation, Medical Image Analysis, Big Data**

I. Introduction

Advances in deep learning (DL) are ushering forward Artificial Intelligence (AI)-driven solutions in industries such as manufacturing, finance and healthcare. Deep neural networks (DNNs) are favored over traditional machine learning (ML) approaches in data-intensive tasks due to their ability to more accurately learn complex patterns from vast amounts of raw data [1]. Thus, DL techniques are widely used in applications with large quantities of high-dimensional data, such as natural language processing, robotics, and computer vision. One of the most impactful applications of DL, specifically computer vision, is medical image analysis (MedIA) where disease diagnosis accuracy in tasks such as X-ray analysis and tumor detection is critical.

It has been shown that DL models can achieve similar or even greater performance than expert analysts in many MedIA applications such as retinopathy [2], radiology [3], and pathology [4]. The success of convolutional neural network (CNN) models, which are DNN models created for image analysis [5], in diagnosis and classification tasks due to their ability to learn complex patterns from raw data creates opportunities for DL technologies to automate certain medical tasks [6].

DL models and CNNs are particularly applicable when working with "Big Data" in MedIA [7] [8] since they can learn high-level features across large amounts of data leading to more accurate results—a vital metric in disease diagnosis. One important consideration for DL systems is the privacy of the confidential and sensitive patient data used to train such models. Due to the security risks that come with aggregating all data into one centralized dataset, privacy-preserving Big Data systems for AI-automated MedIA can be created through a federated learning system on an edge computing architecture which would allow local models on local client devices to train on their own data and send the weights back to a centralized server for aggregation to the global model [9] [10]. Edge computing [11] is a distributed computing system that runs computations and data storage closer to the sources of data, or the "edge" of the network, instead of on a centralized server. Federated learning (FL) [12] is an extension of ML on the edge where local models are trained on local data sources and their parameters are aggregated on a centralized global server without the necessity for centralized data storage. Together, FL and edge computing can lay the foundation of privacy preserving DL systems for MedIA on Big Data.

However, the largest threat to developing such systems is the existence of adversarial attacks. DL models, especially CNNs, have been shown to be extremely vulnerable to adversarial attacks in the form of subtle, carefully engineered perturbations, or calculated 'noise', added to the input data during model training or predictions [13]. Such perturbed data, termed "adversarial samples" or "adversarial examples", can cause misclassification by the target model leading to a significant overall decrease in model accuracy [14]. An example of a perturbed chest X-ray scan for increasing perturbation strength and its resulting effect on the model's classification is shown in Fig. 1. The model correctly classifies the clean image (top left) as showing signs of pneumonia with a 98% accuracy. However, when the image is perturbed with increasing perturbation strength, shown by increasing values of the scale factor ϵ, the model's classification accuracy decreases until it incorrectly

The work was supported in part by the U.S. National Science Foundation under grant NSF-2348417.

classifies the chest X-ray as 'normal' with a confidence of 91% (bottom right).

Furthermore, while FL on edge computing systems preserves the privacy of local data by training on the edge, its local models are still susceptible to adversarial perturbations which can poison the overarching global model, in turn affecting all other local models [15]. Adversarial attacks pose a significant threat to DL systems in MedIA since the high standardization and quality control of medical image data leads to a higher susceptibility to perturbations [16]. Thus, even small adversarial perturbations on clean images can significantly distort model performance and cause catastrophic misdiagnosis [17].

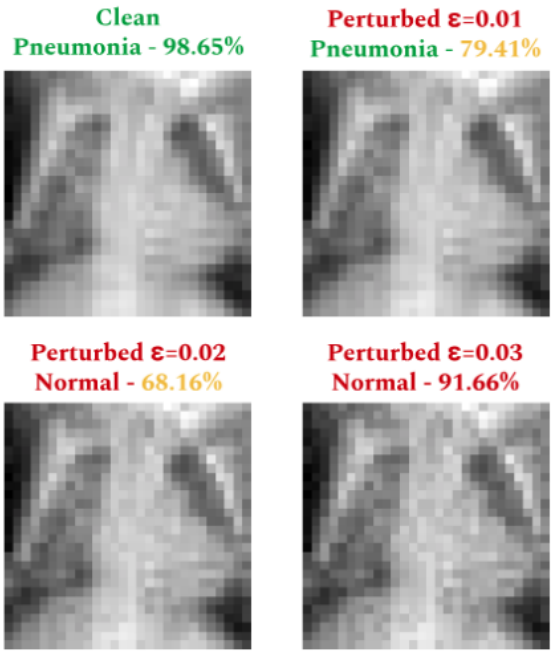

Fig. 1. Adversarial perturbation on chest X-ray scan

This study is an effort to improve on the robustness of DL models against adversarial attacks in the MedIA field and the usability and scalability of defenses against such attacks in FL edge computing systems using Big Data. Our salient contributions are listed as follows.

- First, we utilize transfer learning by training three well-known state-of-the-art pre-trained CNNs such as DenseNet121, MobileNet-V2, and ResNet50 on several MedIA and disease diagnosis datasets of different sizes and evaluate our models' accuracies and robustness against the Fast Gradient Sign Method (FGSM) adversarial attack [14].
- Second, we apply the adversarial training and defensive distillation techniques to improve the robustness of these models and show how they can be used for edge computing and scaled for MedIA with Big Data.

The goal of this research is to advance the field by enabling more secure, accurate, and scalable AI-powered disease diagnosis systems.

The rest of the paper is organized as follows. In Section II, we summarize previously proposed adversarial attacks and defenses. In Section III, we outline the transfer learning models, adversarial attack, and defensive techniques we utilize and evaluate. In Section IV, we detail the evaluation performance of our models and defenses and draw conclusive remarks in Section V.

II. Related Work

In this section, we summarize the previous research relevant to this study done in adversarial attacks and defenses in the fields of DL, FL, and MedIA. In Section II-A, we explore the taxonomy of adversarial attacks and describe several previously established perturbation generation techniques. In Section II-B, we outline the taxonomy of adversarial defenses and summarize established defensive techniques.

A. Adversarial Attacks

We define two categories of adversarial attacks based on adversarial capabilities: black-box attacks, where an adversary assumes no knowledge about the model, and white-box attacks, where the adversary possesses full knowledge about the model's architecture and parameters [18]. There are three broad types of adversarial attacks scenarios: evasion, poisoning, and exploratory attacks. Evasion attacks occur when an adversary has white-box access to the model and attempts to force the model to misclassify an input by crafting adversarial samples during the prediction phase. On the other hand, poisoning attacks are orchestrated during the training phase when an adversary with white-box access attempts to inject adversarial samples to contaminate the training data and compromise the training process, poisoning the model. As opposed to evasion and poisoning attacks where an adversary has information about the model's architecture and weights, exploratory attacks are black-box attacks where an adversary creates surrogate models to approximate and exploit as much knowledge as possible from the target model and its training data.

While FL is susceptible to the same types of adversarial attacks as traditional DL, FL models face different threats at each phase of its execution: Data and Behavior Auditing, Training, and Prediction [15]. In the Data and Behavior Auditing phase, where local workers on the edge send their data to their corresponding local model, FL is vulnerable to poisoning attacks where a local client's data or behavior is compromised, affecting the subsequent training of that model. In the Training phase, where local models are trained on local data and the weights are then aggregated server-side for the global model, a malicious or compromised local client could manipulate their training data or model parameters, corrupting the global training process and poisoning the global model. Finally, in the Prediction phase where the trained global model with the aggregated weights updates each local model, the local models are highly susceptible to evasion attacks.

To carry out evasion or poisoning attacks, adversarial examples must be generated. There are many previously established

techniques and algorithms to generate perturbations that have been proven to cause DNNs and CNNs to misclassify. Szegedy et al. [13] first showed the existence of small perturbations that could fool DL models to misclassify by defining the generation of small but effective perturbations as a minimization problem and using a box-constrained Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization algorithm to approximate a solution. To solve the computationally expensive requirements of the L-BFGS approach, Goodfellow et al. [14] developed a faster method to efficiently compute an adversarial perturbation for an image using the gradient of the model's loss function which was termed the Fast Gradient Sign Method (FGSM). Their work was expanded on by Kurakin et al. [19] to create two variants of the FGSM, the Target Class Method which makes the model misclassify to a specific class, and the Basic Iterative Method, sometimes referred to as Projected Gradient Descent (PGD), that directly extends the FGSM to generate adversarial examples iteratively instead of in a one-shot manner. Interestingly, Papernot et al. [20] created the Jacobian-Based Saliency Map Attack (JSMA) which, by computing a saliency map using the gradients of the DNN layer outputs, iteratively perturbs the one most effective pixel at a time until the model is successfully fooled instead of perturbing the entire image. In a similar vein, Su et al. [21] proposed only perturbing a single pixel in the image by using an evolutionary strategy to find the last surviving 'child' perturbation and using it to alter the most effective pixel. Moosavi-Dezfooli et al. devised DeepFool [22], another iterative approach to adversarial attacks that accumulates small perturbations added to the image each iteration to compute the final perturbation once the perturbation sum is enough for the model to misclassify. DeepFool was shown to compute perturbations that are smaller than the ones computed by FGSM while having a similar effectiveness. In the wake of stronger defenses created against formerly proposed adversarial attacks, Carlini and Wagner [23] introduced a set of three stronger optimization-based adversarial attacks. Finally, whereas the previously discussed attacks compute perturbations to force a model to misclassify on a single image, the 'universal' adversarial perturbations computed by Moosavi-Dezfooli et al. [24] were shown to be able to fool a network on almost any image.

There have been several studies that have explored the impact such adversarial attacks can have on MedIA diagnosis and classification systems. Adversarial attacks have been shown to force DL systems to misclassify in both black-box and white-box scenarios for MedIA data modalities including retinopathy [16] [25] [26], chest X-rays [16] [25] [27], dermoscopy [16] [25] [26] [28], and brain MRI scans [28]. In particular, Taghanaki et al. [27] showed the extreme vulnerability of two state-of-the-art CNNs, Inception-ResNet-v2 and Nasnet-Large, to adversarial attacks in chest X-ray image classification. The gradient-based attacks they implemented, FGSM, PGD, DeepFool, BIM, and L-BFGS, were almost completely successful in fooling both networks in both white-box and black-box scenarios. The prior research conducted on the vulnerability of CNNs to adversarial attacks in general, as well as of that applied to MedIA data, highlights the importance of adequate defense measures to combat such attacks.

B. Defensive Techniques

There are three approaches to defending against adversarial attacks as defined in the literature: using a modified training procedure or modified input during prediction, modifying the network architecture, and using external models as network add-ons [29]. The defensive techniques under these three approaches can be further divided into complete defenses, where the objective is to have the network classify the perturbed image correctly, and detection-only defenses, where the objective is to detect and reject adversarial examples. Since the goal of this study is to improve on the robustness and accuracy of the models under adversarial attacks on their data, we will limit our discussion of related works to the complete defenses. Further, it must be noted that since FL models are vulnerable to the same attacks as DL models—just at different phases in the FL process—the defenses proposed against DNNs will also work in a FL setting on an edge computing system when modifying the training process, data input, or model architecture.

The most commonly proposed and utilized first line of defense against adversarial attacks is adversarial training (AT) [13], where adversarial samples are included in the model's training set. AT results in the regularization of the model to reduce overfitting which improves the model's robustness to adversarial perturbations. However, it should be noted that Moosavi-Dezfooli et al. [24] showed that effective adversarial perturbations could be generated again for networks that underwent AT. A prominent modified-input defense is when Dziugaite et al. [30] demonstrated that JPG compression could reverse the drop in classification accuracy due to perturbations generated by the FGSM, which was further supported by studies around using JPG compression to combat the effectiveness of perturbations [31] and counter attacks by FGSM and DeepFool [32].

Gu and Rigazio [33] introduced Deep Contractive Networks and demonstrated that the use of autoencoders improved the robustness of DL models against the L-BFGS attack. Another modified-network defense is gradient regularization or gradient masking in which large variations in the output of a DNN with respect to small changes in its input were penalized, which, when used with adversarial training, was shown to be effective against improving robustness against the L-BFGS, FGSM, and JSMA [34] [35]. The most popular modified-network defense, however, is employing knowledge distillation [36] to use the knowledge of the network, in the form of class probability vectors for the training data, to train itself, improving its robustness to adversarial perturbations [37] [38].

Finally, there are two primary complete network add-on defenses proposed in the literature. Akhtar et al. [39] proposed the adding of pre-input layers, termed Perturbation Rectifying Network, trained to correct an image modified by universal adversarial perturbations [24] so that the model's classification

on the adversarial sample is the same as that on the clean image. Lee et al. [40] used a Generative Adversarial Network (GAN) to train a model that is robust to FGSM-like attacks by generating perturbations for that model while the model tries to correctly classify both clean and perturbed images. In another GAN-based defense, the generator network was used to correct a perturbed image [41].

While there is substantial literature on defending DNNs from adversarial attacks on natural images, research on the effectiveness of proposed complete defenses on MedIA data is limited since due to the challenges for adversarial defenses in MedIA, recent works have focused primarily on detecting and rejecting adversarial samples rather than classifying them correctly [16]. This study aims to contribute to making DL systems for MedIA more robust to adversarial attacks and perturbations rather than just using detection-only approaches.

III. Methods

This study thoroughly assessed the effectiveness of CNN models that have already been trained applied to MedIA. DenseNet121, MobileNet-V2, and ResNet50 were among the three pre-trained models used for transfer learning. Four established MedIA datasets from the MedMNIST2D collection [42] [43] including BreastMNIST [44], PneumoniaMNIST [45], DermaMNIST [46] [47], and OCTMNIST [45] were used in this study. Each dataset was then split into train, validation, and testing sets and preprocessed. Then, all three CNNs were trained on each of the four datasets and evaluated by accuracy on clean and adversarially perturbed testing data. Finally, the models were defended using AT and knowledge distillation and once again evaluated on the clean and perturbed data. To implement the experiment, the study uses Python 3 with the TensorFlow and Keras libraries for training, attacking, and defending the DL models on a local Jupyter notebook environment, along with a Core i7 processor, 16 GB of RAM, and Iris Xe Graphics.

A. Transfer Learning

The ML technique of transfer learning involves using a pre-trained model's knowledge from a prior task and adapting it for an unfamiliar but still related task to avoid starting from scratch [48]. CNNs process images by relying on their convolutional layers to extract high-level features that are used for classification by applying filters, such as edge and corner detectors, to produce convolved copies and feature maps of the original images [49]. Thus, fine-tuning a pre-trained CNN involves first using the general features extracted by the model's initial layers and then optimizing the last layers for the new task [50]. This study used three pre-trained models including MobileNet-V2 [51], a smaller model created for mobile and low-resource environments such as those in FL and edge computing systems, as well as two larger models, DenseNet121 [52] and ResNet50 [53], for their ability to handle and learn patterns from Big Data.

1) DenseNet121: DenseNet121 [52] utilizes a dense layer connectivity pattern where each layer in the network is directly connected to all following layers, creating dense 'blocks' of connections that allow for feature reuse and propagation across layers. Its architecture is made up of 121 layers divided into four dense blocks with transition layers in between them to reduce the dimensionality of features before passing them through to the next block. Each dense block contains multiple 1×1 and 3×3 convolutional layers, where the 1×1 layer serves as a 'bottleneck' layer to decrease computational complexity by reducing the number of input features to the more expensive 3×3 convolutions and only preserving important features. The entire network begins with a larger, 7×7 convolution layer and a max pooling layer to extract important preliminary features and ends with a global average pooling layer and a fully connected layer to simplify and produce a classification output from the final feature maps. Since each layer receives feature maps from all prior layers, this enables deep supervision throughout the network. DenseNet121's unique dense connectivity pattern results in it only having eight million parameters despite its large depth, making it an effective and memory-efficient CNN.

2) MobileNet-V2: MobileNet-V2 [51] is a CNN architecture designed for mobile and computational-resource-constrained settings. It uses inverted residual blocks and linear bottlenecks to improve model performance while maintaining a low resource cost. It consists of a beginning convolutional layer with 32 filters, followed by 19 residual bottleneck layers. Each inverted residual block contains a 1×1 convolutional expansion layer to increase the number of dimensions and features, a 3×3 depthwise convolutional layer to efficiently filter significant features, and a 1×1 projection convolutional layer to project the filtered features back down to a lower dimension to linearly bottleneck complexity. The network ends with a 1×1 convolution, global average pooling, and a fully connected layer to simplify the extracted features and generate a classification output. MobileNet-V2's unique use of linear bottlenecks instead of ReLU helps retain important information in fewer dimensions.

3) ResNet50: The usage of residual learning was introduced in ResNet50 [53], whose architecture has 50 layers, including 48 convolutional layers organized into 4 stages. Each stage of convolutional layers contains multiple residual blocks, where each block has a skip connection that allows the network to learn what features to extract at each layer from that layer's inputs. The network starts with a large 7×7 convolution and a max pooling layer and ends with a global average pooling layer and a fully connected layer for simplifying and using the extracted features to produce a classification output. ResNet's residual connections make it easier to optimize larger and deeper networks since it is easier for the network to learn small changes at a time.

B. Adversarial Attacks

Evasion attacks are the most common type of adversarial attack and are the easiest for an adversary to implement as

they do not require access to the training data like poisoning attacks do [18]. Additionally, white-box attacks, in which an adversary has complete access to the model's architecture and parameters, are the most effective types of attacks on DL models as they have been proven to cause complete misdiagnosis in MedIA [27]. Thus, this study will limit its scope to defending against white-box evasion attacks. Specifically, it will utilize the FGSM [14] to generate adversarial perturbations since it is the easiest method for an adversary to implement, is computationally fast due to its one-shot approach, and has been shown to be very effective against both DL and MedIA networks [27]. The FGSM can efficiently compute an adversarial perturbation for a given image using the following formula:

$$adv\_x = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)), \tag{1}$$

where $\nabla_x J$ calculates the gradient of the loss function J evaluated at the current value of the model parameters, $\theta, x,$ and y, sign() is the sign function, and ϵ is the scale factor, a small scalar value that controls the strength and perceptibility of the perturbation. In this study, a range of values for the scale factor from 0.001–0.005 was used to evaluate the networks' robustness against both weaker and stronger adversarial perturbations.

C. Adversarial Defenses

To defend against the FGSM attack, the two chosen defenses for this study were Adversarial Training (AT) [13] and Knowledge Distillation [37]. These defenses where chosen due to their variety in defensive approach, as AT is a modified-training defense while Knowledge Distillation modifies the network, and their proven effectiveness in defending against the FGSM attack on DL models trained on natural images [13] [38].

AT includes the incorporation of adversarial samples into the network's training data to regularize it and reduce its overfitting on clean data, which helps the model generalize to classify adversarial samples correctly. The FGSM was used with a scale factor of $\epsilon = 0.001$ to generate weaker adversarial samples for AT. This defense was implemented by further training the transfer learning CNNs on adversarially perturbed data to improve their robustness.

Knowledge Distillation transfers the knowledge of a larger, more complex 'teacher' network to a smaller 'student' network [36]. When using distillation as a defense, the knowledge of the teacher network is extracted as class probability vectors of the model's training data and is fed back to train the student model. To implement this, the DenseNet121 models trained for multi-class classification on the DermaMNIST and OCTMNIST datasets were used as the teacher models and a new, smaller custom model was used as the student model and was trained using distillation with a Kullback-Leibler Divergence loss function.

After the networks were defended using AT and Knowledge Distillation, their robustness was once again evaluated against adversarial perturbations generated by the FGSM.

IV. Experimental Results

In this section, we present and analyze our evaluation results of the performance of the three transfer learning CNNs on clean and FGSM-perturbed data for all four datasets, as well as the effectiveness of AT and Knowledge Distillation as defensive techniques.

A. Data and Preprocessing

The MedMNIST collection [42] [43] consists of 10 standardized datasets from various medical data modalities including X-ray, OCT, ultrasounds, and CT scans. Its datasets vary in scale, from datasets with a few hundred or thousand samples to those with over 200,000. Furthermore, all datasets are already standardized by the authors by pre-processing and splitting the data into training-validation-testing subsets, using the data split from the source datasets if provided, or using a 7:1:2 (train:val:test) split otherwise. In this study we utilized both small and large datasets to demonstrate the scalability of adversarial defenses in MedIA with Big Data.

1) BreastMNIST: BreastMNIST, the smallest dataset used in this study, is based on a dataset [44] of 780 breast ultrasound images initially divided into 3 classes of normal, benign, and malignant but modified by the authors to be suited for binary classification between 'normal/benign' and 'malignant'. It uses a 7:1:2 split ratio for training, validation, and testing sets and contains images resized to $1 \times 28 \times 28$.

2) PneumoniaMNIST: PneumoniaMNIST is based on a prior dataset [45] of 5,856 pediatric chest X-ray scans divided into 2 classes of 'pneumonia' and 'normal' for binary classification. The original training set was split into a 9:1 ratio for training and validation and the source validation set was used as the testing set. The images are single-channel and were resized by the authors to $1 \times 28 \times 28$.

3) DermaMNIST: DermaMNIST is based on the HAM10000 dataset [46] which contains 10,015 dermatoscope images of common pigmented skin lesions. The images are labeled into 7 classes representing 7 different skin diseases, lending itself to a multi-class classification task. The dataset uses a 7:1:2 training-validation-testing split and the original images were resized by the authors to $3 \times 28 \times 28$.

4) OCTMNIST: OCTMNIST, the largest dataset used in this study, is based on a dataset [45] of 109,309 optical coherence topography (OCT) scans for retinal diseases. The images are labeled into 4 types of retinal diseases, lending the dataset to a multi-class classification task. The source training set was split into a 9:1 ratio for training and validation sets and the source validation set was used as the testing set. The images are single-channel and were resized by the authors to $1 \times 28 \times 28$.

For all the datasets used in this study, the MedMNIST-proposed train-validation-test split was used. However, the datasets underwent preprocessing before transfer learning with pre-trained CNNs. The datasets containing single-channel images, namely BreastMNIST, PneumoniaMNIST, and OCTMNIST, had their images' dimensions expanded and channels replicated to transform them from dimensions of $1 \times 28 \times 28$ to

3-channel images of dimensions $3 \times 28 \times 28$ to fit the requirements of the CNNs used for transfer learning. Furthermore, all datasets were normalized by scaling their images' pixels down by a factor of 255 to the range of 0 to 1 and resizing the images from $3 \times 28 \times 28$ to $3 \times 32 \times 32$ to be valid inputs for the pre-trained CNNs.

B. Transfer Learning Performance

After all datasets were preprocessed, each network was trained over each dataset for 5 epochs with a batch size of 32, using the Adam optimization method and either the Binary Cross entropy loss function or Sparse Categorical Cross entropy loss function, depending on whether the task was binary or multi-class classification. Then, the networks were evaluated for accuracy on the testing datasets. The evaluation results are shown in Table I.

TABLE I
TRANSFER LEARNING EVALUATION RESULTS

Evaluation Accuracy	**Model**		
Dataset	***DenseNet121***	***MobileNetV2***	***ResNet50***
BreastMNIST	83.97%	78.85%	N/A*
PneumoniaMNIST	88.62%	84.29%	86.70%
DermaMNIST	72.82%	68.08%	71.77%
OCTMNIST	77.78%	69.25%	75.56%

*ResNet50 evaluation results on BreastMNIST excluded

When training ResNet50 on the BreastMNIST dataset, it yielded an extremely low evaluation accuracy of a mere 25%. We diagnosed this phenomenon to be a result of the mismatch between ResNet50's large number of parameters and the BreastMNIST dataset's small number of only 780 samples, which was not suitable for such a large model. This mismatch led to the model completely overfitting to the training dataset resulting in a poor generalization to the testing set. Furthermore, we found the BreastMNIST dataset to be highly unbalanced, as out of the 546 samples in the training set, 399 were 'normal/benign' whereas only 147 samples were 'malignant'. To attempt to fix the small sample size issue, we attempted to upsample the dataset using data augmentation techniques including rotations, shifts, shears, flips, and zooms of the images to synthesize more data for training. Additionally, we undersampled the abundant class to attempt to balance the dataset. However, our synthetic images were not realistic-enough variations and did not improve ResNet50's performance on the data. This raises an important insight when choosing network architectures for smaller tasks since the model size and complexity must match the difficulty of the task and size of the dataset. Since ResNet50 was chosen to show the scalability of CNN classifiers for Big Data not smaller datasets, and it had a consistently poor performance on the BreastMNIST dataset, the results of training ResNet50 on BreastMNIST were excluded from this study and it was not further attacked or defended to maintain the reliability and validity of this study's results.

With ResNet50 on BreastMNIST being the only exception, all the models trained in this study achieved an evaluation accuracy close to the official MedMNIST benchmarking accuracy [42]. It was discovered that detecting pneumonia from the PneumoniaMNIST dataset was one of the easier tasks that could be done with a higher accuracy of around 85%, while classifying skin diseases from the DermaMNIST dataset was a much harder task, having a lower benchmark and evaluation accuracy around 70%. DenseNet121 was shown to consistently outperform MobileNet-V2 and ResNet50 on all the datasets, highlighting the effectiveness of its densely connected structure in providing deep supervision throughout the network during the learning process. ResNet50 was close behind, its large number of parameters and residual block learning architecture allowing it to pick up on complex patterns from the data. DenseNet121's depth and ResNet50's high number of parameters makes these models highly applicable to MedIA scenarios with Big Data, shown by their outstanding performance for the large OCTMNIST dataset that aligns with the dataset's benchmark [42]. Finally, MobileNet-V2 slightly underperformed the other networks on all datasets due to its much smaller depth and number of parameters. However, its good performance despite its small size and number of parameters makes it an effective and cost-efficient choice for FL scenarios on mobile devices on the edge.

C. Adversarial Attack Evaluation

After the transfer learning models were trained and evaluated on all the datasets, perturbed testing sets were generated for each dataset and model pair and the models were evaluated on those perturbed test sets. The adversarial examples were generated using the FGSM attack and 5 perturbed test sets were generated for each model ranging from weaker perturbations of $\epsilon = 0.001$ to stronger perturbations of $\epsilon = 0.005$. The transfer learning models were evaluated on these 5 perturbed test datasets and their evaluation results are shown in Table II.

All models experienced decreases in performance and accuracy as a result of the FGSM attack, and the accuracy drops were larger for stronger perturbations notated by increasing scale factors ϵ. Across the datasets, ResNet50 usually had a higher inbuilt robustness to the FGSM perturbations and experienced a smaller drop in performance than DenseNet121 and MobileNetV2 likely due to its residual block architecture which allowed the network to learn feature transformations in smaller steps at each layer. Surprisingly, ResNet50 kept an evaluation accuracy of greater than 70% on the PneumoniaMNIST data even when perturbed with a scale factor of $\epsilon = 0.005$. On the other hand, due to DenseNet121's and MobileNetV2's smaller size and number of parameters they were more susceptible to the perturbations leading to larger drops in accuracy. These results show the devastating effect the FGSM attack can have on DL systems for MedIA since DenseNet121's earlier superior performance to the other models on all data and applicability when dealing with Big Data was replaced by its inferior performance on the adversarially perturbed data. Additionally, MobileNet-V2, the most usable model for FL systems due to its high performance despite its small size, also experienced a significant degradation in

TABLE II
MODEL PERFORMANCE ON CLEAN AND ADVERSARIALLY PERTURBED DATA

Dataset	Model	Clean	$\epsilon = 0.001$	$\epsilon = 0.002$	$\epsilon = 0.003$	$\epsilon = 0.004$	$\epsilon = 0.005$
BreastMNIST	DenseNet121	**83.97%**	**71.15%**	56.41%	50.00%	43.59%	38.46%
	MobileNetV2	78.85%	62.82%	**57.05%**	**54.49%**	**48.72%**	**44.87%**
PneumoniaMNIST	DenseNet121	**88.62%**	80.61%	71.96%	66.51%	59.29%	54.01%
	MobileNetV2	84.29%	67.95%	58.65%	54.01%	51.60%	50.64%
	ResNet50	86.70%	**82.69%**	**79.33%**	**76.76%**	**72.92%**	**70.83%**
DermaMNIST	DenseNet121	**72.82%**	65.64%	60.25%	56.41%	53.67%	51.07%
	MobileNetV2	68.08%	62.29%	60.15%	58.50%	57.91%	**58.00%**
	ResNet50	71.77%	**68.33%**	**65.39%**	**62.04%**	**59.20%**	57.11%
OCTMNIST	DenseNet121	**77.78%**	65.52%	55.21%	46.39%	39.49%	34.40%
	MobileNetV2	69.25%	52.21%	42.16%	36.26%	32.27%	29.54%
	ResNet50	75.56%	**69.55%**	**62.61%**	**55.96%**	**48.74%**	**41.74%**

TABLE III
COMPARISON OF ORIGINAL, ADVERSARIALLY TRAINED (AT), AND KNOWLEDGE DISTILLATION (KD) MODEL PERFORMANCE

Dataset	Model	Defense	Clean	$\epsilon = 0.001$	$\epsilon = 0.002$	$\epsilon = 0.003$	$\epsilon = 0.004$	$\epsilon = 0.005$
BreastMNIST	DenseNet121	Original	**83.97%**	71.15%	56.41%	50.00%	43.59%	38.46%
		AT	78.85%	**75.00%**	71.15%	67.31%	61.54%	58.97%
	MobileNetV2	Original	78.85%	62.82%	57.05%	54.49%	48.72%	44.87%
		AT	71.79%	73.08%	**75.00%**	**75.00%**	**74.36%**	**74.36%**
PneumoniaMNIST	DenseNet121	Original	**88.62%**	80.61%	71.96%	66.51%	59.29%	54.01%
		AT	86.86%	81.41%	77.88%	71.96%	67.63%	64.10%
	MobileNetV2	Original	84.29%	67.95%	58.65%	54.01%	51.60%	50.64%
		AT	77.08%	74.68%	70.67%	67.79%	65.06%	62.02%
	ResNet50	Original	86.70%	82.69%	79.33%	76.76%	72.92%	70.83%
		AT	84.94%	**83.01%**	**80.77%**	**78.85%**	**76.44%**	**74.84%**
DermaMNIST	DenseNet121	Original	72.82%	65.64%	60.25%	56.41%	53.67%	51.07%
		AT	**72.92%**	70.02%	68.78%	67.58%	65.59%	63.44%
		KD	68.98%	68.98%	68.93%	68.83%	**68.83%**	**68.73%**
	MobileNetV2	Original	68.08%	62.29%	60.15%	58.50%	57.91%	58.00%
		AT	65.54%	67.03%	66.98%	67.13%	66.78%	66.33%
	ResNet50	Original	71.77%	68.33%	65.39%	62.04%	59.20%	57.11%
		AT	70.97%	**70.12%**	**69.58%**	**69.03%**	68.53%	67.83%
OCTMNIST	DenseNet121	Original	77.78%	65.52%	55.21%	46.39%	39.49%	34.40%
		AT	75.29%	71.79%	68.45%	64.98%	61.84%	58.07%
		KD	**78.69%**	**78.36%**	**78.16%**	**77.81%**	**77.46%**	**77.13%**
	MobileNetV2	Original	69.25%	52.21%	42.16%	36.26%	32.27%	29.54%
		AT	61.79%	61.48%	58.87%	55.09%	51.37%	48.50%
	ResNet50	Original	75.56%	69.55%	62.61%	55.96%	48.74%	41.74%
		AT	–	–	–	–	–	–

performance. The evaluation results of the models under the FGSM perturbations underscore the importance and necessity of defensive measures to improve the networks' robustness to adversarial attacks.

D. Defense Performance Evaluation

Once the transfer learning models were evaluated against the FGSM perturbations, we employed adversarial training and trained knowledge distillation models to defend against the adversarial attack. The adversarially-trained and knowledge distillation models were then evaluated on the clean and perturbed data and their evaluation results are shown in Table III.

Adversarial training was an effective defense measure as it improved all models' robustness across all the datasets and reversed drops in accuracy due to the perturbed data. AT had a more prominent effect against stronger perturbations with higher values of ϵ. One negative side effect of AT was that it led to a small decrease in the adversarially-trained model's accuracy on the clean data since training it to classify adversarial examples correctly shifts the class boundaries in the network's latent space, causing it to occasionally classify a clean image incorrectly. However, considering the significant improvement in accuracy on the perturbed data, this small drop in accuracy on the clean data is a justifiable trade-off. Furthermore, while MobileNet-V2 had the lowest inbuilt robustness due to its smaller size and number of parameters, it showed the greatest improvement under AT especially for the BreastMNIST and DermaMNIST datasets. While AT was very effective in improving the networks' accuracies on the perturbed data for the smaller datasets, it did not have as

prominent of an effect on the larger dataset, OCTMNIST. Furthermore, due to the hardware limitations of this study, we were unable to adversarially-train ResNet50 on OCTMNIST due to its complexity and large number of parameters.

Since multi-class classification is a more complex task than binary classification and is more common in the field of MedIA, we only tested the knowledge distillation defense against perturbations on the DermaMNIST and OCTMNIST datasets. Additionally, due to the computationally heavy requirements of AT for large datasets like OCTMNIST and Big Data in general, knowledge distillation was tested with DenseNet121 as its deep connectivity structure and large depth but small number of parameters can handle large amounts of data. However, the hardware limitations and resource constraints of this study prevented knowledge distillation from being used with MobileNet-V2 and ResNet50. For DermaMNIST, knowledge distillation with DenseNet121 outperformed both the DenseNet121 teacher model and all AT models for the stronger perturbations but slightly underperformed for the clean data and weaker perturbations. For OCTMNIST, however, the distilled model achieved greater results and significantly higher accuracies than all other models tested on OCTMNIST for the clean data and all perturbed test data. This outcome demonstrates that using knowledge distillation as a defense is more effective for larger datasets and Big Data, especially in FL scenarios, since a model like DenseNet121 can effectively teach a smaller CNN to extract and learn complex features from vast amounts of data, and the smaller network can be used on devices on the edge.

V. Conclusion

The main objective of this study was to improve on the robustness of DL models against adversarial attacks in the MedIA field while ensuring scalability for Big Data applications in FL and edge computing systems. Three pre-trained transfer learning models of varying depths and parameter counts, DenseNet121, MobileNet-V2, and ResNet50, were fine-tuned on four MedIA datasets of varying sizes, BreastMNIST, PneumoniaMNIST, DermaMNIST, and OCTMNIST, and evaluated for accuracy. All three models achieved high accuracy corresponding to the benchmark for each dataset, even able to handle large datasets like OCTMNIST, with DenseNet121 outperformance of its counterparts making it a prime candidate for large-scale applications.

When adversarially attacked with the FGSM, however, all three models showcased an extreme vulnerability to the generated perturbations leading to significant drops in accuracy, with ResNet50 showing the highest inbuilt robustness and DenseNet121 showing one of the lowest. Using AT to defend the models against the FGSM attack proved effective in improving model robustness against the perturbations across all the datasets, especially for the smaller ones, although it came with a slight tradeoff in model performance on the clean data. Knowledge distillation was shown to be highly effective in reversing accuracy drops due to perturbations on large datasets like OCTMNIST, offering a more scalable solution for Big Data scenarios in FL and edge computing.

Robust and scalable solutions are essential for accurate disease diagnosis DL systems, especially when working with Big Data and privacy-preserving architectures. While adversarial attacks remain a significant threat to such systems, the application of defensive techniques such as AT and knowledge distillation can mitigate security risks. Future research should build on the results of this study by testing the application of other pre-trained CNNs or customized architectures to determine whether they can enhance classification performance on MedIA data. Furthermore, the usage of stronger adversarial attacks and the evaluation and construction of more advanced defensive techniques should be applied to the MedIA field. Finally, future work should focus on deploying these models and defensive strategies within a FL and edge computing environment to assess the robustness and security of such environments at scale.

References

[1] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, Mar. 2021. [Online]. Available: http://dx.doi.org/10.1186/s40537-021-00444-8

[2] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, p. 2402, Dec. 2016. [Online]. Available: http://dx.doi.org/10.1001/jama.2016.17216

[3] K. Murphy, S. S. Habib, S. M. A. Zaidi, S. Khowaja, A. Khan, J. Melendez, E. T. Scholten, F. Amad, S. Schalekamp, M. Verhagen, R. H. H. M. Philipsen, A. Meijers, and B. van Ginneken, "Computer aided detection of tuberculosis on chest radiographs: An evaluation of the cad4tb v6 system," *Scientific Reports*, vol. 10, no. 1, Mar. 2020. [Online]. Available: http://dx.doi.org/10.1038/s41598-020-62148-y

[4] R. Aggarwal, V. Sounderajah, G. Martin, D. S. W. Ting, A. Karthikesalingam, D. King, H. Ashrafian, and A. Darzi, "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis," *npj Digital Medicine*, vol. 4, no. 1, Apr. 2021. [Online]. Available: http://dx.doi.org/10.1038/s41746-021-00438-z

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.

[6] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices," *npj Digital Medicine*, vol. 1, no. 1, Aug. 2018. [Online]. Available: http://dx.doi.org/10.1038/s41746-018-0040-6

[7] A. Tahmassebi, A. Ehtemami, B. Mohebali, A. H. Gandomi, K. Pinker, and A. Meyer-Baese, "Big data analytics in medical imaging using deep learning," in *Big Data: Learning, Analytics, and Applications*, F. Ahmad, Ed. SPIE, May 2019, p. 13. [Online]. Available: http://dx.doi.org/10.1117/12.2516014

[8] S. Bagga, S. Gupta, and D. K. Sharma, *Big Data analytics in medical imaging*. Elsevier, 2021, p. 113–136. [Online]. Available: http://dx.doi.org/10.1016/B978-0-12-820203-6.00006-0

[9] P. Pace, G. Aloi, R. Gravina, G. Caliciuri, G. Fortino, and A. Liotta, "An edge-based architecture to support efficient applications for healthcare industry 4.0," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, p. 481–489, Jan. 2019. [Online]. Available: http://dx.doi.org/10.1109/TII.2018.2843169

[10] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, "Federated learning for medical image analysis: A survey," 2023. [Online]. Available: https://arxiv.org/abs/2306.05980

[11] K. Cao, Y. Liu, G. Meng, and Q. Sun, "An overview on edge computing research," *IEEE Access*, vol. 8, p. 85714–85728, 2020. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2020.2991734

[12] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," 2016. [Online]. Available: https://arxiv.org/abs/1602.05629

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2014. [Online]. Available: https://arxiv.org/abs/1312.6199

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[15] P. Liu, X. Xu, and W. Wang, "Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives," *Cybersecurity*, vol. 5, no. 1, Feb. 2022. [Online]. Available: http://dx.doi.org/10.1186/s42400-021-00105-6

[16] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 110, p. 107332, Feb. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2020.107332

[17] G. Bortsova, C. González-Gonzalo, S. C. Wetstein, F. Dubost, I. Katramados, L. Hogeweg, B. Liefers, B. van Ginneken, J. P. Pluim, M. Veta, C. I. Sánchez, and M. de Bruijne, "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," *Medical Image Analysis*, vol. 73, p. 102141, Oct. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.media.2021.102141

[18] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," 2016. [Online]. Available: https://arxiv.org/abs/1611.03814

[19] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2017. [Online]. Available: https://arxiv.org/abs/1611.01236

[20] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," 2015. [Online]. Available: https://arxiv.org/abs/1511.07528

[21] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, p. 828–841, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1109/TEVC.2019.2890858

[22] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," 2016. [Online]. Available: https://arxiv.org/abs/1511.04599

[23] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1608.04644

[24] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," 2017. [Online]. Available: https://arxiv.org/abs/1610.08401

[25] S. G. Finlayson, H. W. Chung, I. S. Kohane, and A. L. Beam, "Adversarial attacks against medical deep learning systems," 2019. [Online]. Available: https://arxiv.org/abs/1804.05296

[26] U. Ozbulak, A. Van Messem, and W. De Neve, *Impact of Adversarial Examples on Deep Learning Models for Biomedical Image Segmentation*. Springer International Publishing, 2019, p. 300–308.

[27] S. A. Taghanaki, A. Das, and G. Hamarneh, "Vulnerability analysis of chest x-ray image classification against adversarial attacks," 2018. [Online]. Available: https://arxiv.org/abs/1807.02905

[28] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. robustness: Adversarial examples for medical imaging," 2018. [Online]. Available: https://arxiv.org/abs/1804.00504

[29] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.

[30] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, "A study of the effect of jpg compression on adversarial images," 2016. [Online]. Available: https://arxiv.org/abs/1608.00853

[31] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," 2018. [Online]. Available: https://arxiv.org/abs/1711.00117

[32] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, "Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression," 2017. [Online]. Available: https://arxiv.org/abs/1705.02900

[33] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," 2015. [Online]. Available: https://arxiv.org/abs/1412.5068

[34] A. S. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," 2017. [Online]. Available: https://arxiv.org/abs/1711.09404

[35] C. Lyu, K. Huang, and H.-N. Liang, "A unified gradient regularization family for adversarial examples," 2015. [Online]. Available: https://arxiv.org/abs/1511.06385

[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[37] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," 2016. [Online]. Available: https://arxiv.org/abs/1511.04508

[38] N. Papernot and P. McDaniel, "On the effectiveness of defensive distillation," 2016. [Online]. Available: https://arxiv.org/abs/1607.05113

[39] N. Akhtar, J. Liu, and A. Mian, "Defense against universal adversarial perturbations," 2018. [Online]. Available: https://arxiv.org/abs/1711.05929

[40] H. Lee, S. Han, and J. Lee, "Generative adversarial trainer: Defense to adversarial perturbations with gan," 2023. [Online]. Available: https://arxiv.org/abs/1705.03387

[41] S. Shen, G. Jin, K. Gao, and Y. Zhang, "Ape-gan: Adversarial perturbation elimination with gan," 2017. [Online]. Available: https://arxiv.org/abs/1707.05474

[42] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 191–195.

[43] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[44] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, no. 104863, p. 104863, Feb. 2020.

[45] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. L. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. N. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, and K. Zhang, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, Feb. 2018.

[46] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, p. 180161, Aug. 2018.

[47] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," 2019. [Online]. Available: https://arxiv.org/abs/1902.03368

[48] R. Kaur, R. Kumar, and M. Gupta, "Review on transfer learning for convolutional neural network," in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE, Dec. 2021.

[49] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015. [Online]. Available: https://arxiv.org/abs/1511.08458

[50] C. Iorga and V.-E. Neagoe, "A deep CNN approach with transfer learning for image recognition," in *2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. IEEE, Jun. 2019.

[51] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2019. [Online]. Available: https://arxiv.org/abs/1801.04381

[52] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018. [Online]. Available: https://arxiv.org/abs/1608.06993

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385