Abstract—Early identification of COVID-19 using a deep model trained on Chest X-Ray and CT images has gained considerable attention from researchers to speed up the process of identification of active COVID-19 cases. These deep models act as an aid to hospitals that suffer from the unavailability of specialists or radiologists, specifically in remote areas. Various deep models have been proposed to detect the COVID-19 cases, but few works have been performed to prevent the deep models against adversarial attacks capable of fooling the deep model by using a small perturbation in image pixels. This paper presents an evaluation of the performance of deep COVID-19 models against adversarial attacks. Also, it proposes an efficient yet effective Fuzzy Unique Image Transformation (FUIT) technique that downsamples the image pixels into an interval. The images obtained after the FUIT transformation are further utilized for training the secure deep model that preserves high accuracy of the diagnosis of COVID-19 cases and provides reliable defense against the adversarial attacks. The experiments and results show the proposed model prevents the deep model against the six adversarial attacks and maintains high accuracy to classify the COVID-19 cases from the Chest X-Ray image and CT image Datasets. The results also recommend that a careful inspection is required before practically applying the deep models to diagnose the COVID-19 cases.

Index Terms—Adversarial Attacks, Chest X-Ray, COVID-19, CT Image, Deep Models, Fuzzy Unique Image Transformation.

I. INTRODUCTION AND RELATED WORK

The occurrence of a novel CORONA VIRUS [1] challenges the healthcare systems of all across the world to control an exponential growth of CORONA VIRUS that first occurred in Wuhan and Hebei cities of the China [1] in December 2019 and later spared to other countries across the world. Based on the degree of spread of the virus World Health Organization (WHO) declared the disease as COVID-19 pandemic [2]. Cough, fatigue, fever, and illness in the lungs are among the earlier symptoms suggested by clinical experts for diagnosing COVID-19 cases at an initial stage. Control and prevention of the COVID-19 demand the maximum number of medical tests. Healthcare systems across the world suffer from a lack of effective testing kits to identify COVID-19 cases in a current situation. The early identification of COVID-19 cases would be helpful to quarantine the high-risk COVID-19 patients and also useful to break a chain of further spread of the virus in the community.

In an attempt to develop a testing kit for the diagnosis of COVID-19, researchers from the radiology domain suggested the use of reverse transcription-polymerase chain reaction (RT-PCR) test [3]. However, the test requires long latency to identify COVID-19 cases and demands highly expert radiologists [3]. The RT-PCR test also suffers from a high false-positive rate during the diagnosis of COVID-19 cases [3], which is not acceptable. A good survey of the various image, sound, and blood test report-based datasets available for diagnosing COVID-19 cases can be found in [3]. Recent studies [5], [6] have shown Chest X-Ray images of COVID-19 patients play a vital role in timely identification and further control of the COVID-19 cases. Inspired by the success of work on chest X-Ray and CT scan images, various methods and computer-aided systems have been proposed that combine deep learning methods and radiology expert knowledge to identify the COVID-19 cases. A comprehensive study of various deep learning-based methods for diagnosis of COVID-19 using chest X-Ray and CT Scan images can be found in [7], [8], [9], [10]. The majority of the deep learning models proposed for identifying COVID-19 cases are based on transfer learning [9], [11], [12], [13], attention-based mechanism [13], [15], [16], [17], self-supervised learning [18], [19] and explainable deep models [20], [21], [22], [23]. On the other hand, very little work has been performed towards the vulnerability of deep models against adversarial attacks [24] capable of misleading the deep model with a small perturbation in pixels of an input image. Identification of COVID-19 cases requires expert opinions over the chest X-Ray and CT scan images. It also involves the communication of the COVID-19 data through the web to receive the expert’s suggestions and reports.

The deep learning models have achieved new heights of state-of-the-art (SOTA) methods in object detection [25], text mining [26], speech recognition [27] and computer vision [28]. However, it has been well explored that the deep models are sensitive towards small perturbation in an input and easily fooled by the attacker. This paradigm is also known as Adversarial attack [24], [29]. The study of adversarial attacks was introduced a decade ago [30] and gained huge attention from researchers of deep learning due to the increasing demand for deep learning techniques in various real-life applications. Data and models privacy and security concerns make the study of adversarial attacks popular in deep learning research. The existence of the adversarial attacks put various questions on the generalization of deep models for the diagnosis of COVID-19 using medical images. In [31], Hirano et al. investigated the performance analysis of deep models for the diagnosis of COVID-19 cases in the presence of adversarial attacks. A previous study suggested the vulnerability as a major bottleneck for the medical image-based diagnosis [32].
Adversarial attacks on deep models can be subdivided into two major classes. The first type of attack is known as a white-box attack [30], and the second type of attack is known as a black-box attack [30]. The white-box attacks use the full knowledge of the deep model, dataset, architecture, and parameters. However, the scenario is different in black-box attacks that only partially access the information related to deep models. The proposed method aims to provide defense against white-box attacks that are very hard to prevent in practical scenarios. Adversarial attacks are further broadly classified into two classes, targeted and untargeted attacks. The targeted attacks modify the clean images into adversarial images that make the deep model to classify the input image into a class set by the attackers. For an example, if a clean image of the non-COVID-19 case is transformed into an adversarial image with a target label set as COVID-19 case and the model classifies the images as COVID-19 instead of a normal case. On the other hand, in case of untargeted attacks, the image is transformed into an adversarial image such that the model classifies the image into labels other than the true class label of the image. For an example, the image belongs to the COVID-19 case misclassified as a normal case or pneumonia case after an untargeted adversarial attack. The work presents in this paper intends to present a defense mechanism against an untargeted class of adversarial attacks.

A pioneer work that focuses on the generation of the adversarial examples was presented by Goodfellow et al. [33]. The author proposed a fast gradient-based approach to generate adversarial samples. By taking inspiration from the initial work of Goodfellow et al. [33], various methods have been proposed to generate adversarial examples. Moosavi et al. [34] proposed a deep fool mechanism that generates perturbation until the confidence of the model decreases on the correct label for the given input. The iteration to create perturbation stops when the deep model is fooled. In [35], the author proposed an adversarial mechanism that uses an Adam optimization method [56] for an adversarial attack. Sharma et al. [37] proposed a framework that uses attention feature maps to generate adversarial examples to attack the deep model.

Besides the development of adversarial attack techniques, numerous defense methodologies [38], [39] have been proposed to prevent the deep models against the adversarial attacks. The defense methodologies are further grouped into two categories black box defense and white box defense. The white box defense involves adversarial images as input to train the deep model. The adversarial images are generated by one of the adversarial attack techniques [33], [34] mentioned above. On the contrary, the black box defense does not involve the adversarial images to train the deep model that prevents adversarial attacks. Data augmentation techniques [39], input transformation [40] and an encryption inspired shuffling of images [41] are among the popular techniques that are well explored to perform black-box defense. A comprehensive study of various defense techniques against the adversarial attacks can be found in [38]. The white box defenses are more successful as compared to the black box defense. Still, they suffer from a high probability of failure against the attacks having a complexity greater than the adversarial attacks employed to generate the adversarial images while training the white box defense models. The black box defense is independent of the complexity of the attack mechanism thus gained more attention to develop robust and secure deep models against the adversarial attacks.

The proposed fuzzy unique image transformation (FUIT) technique belongs to the black box defense category. To the best of our knowledge, a fuzzy logic-based black box defense has not been proposed to prevent the COVID-19 images from adversarial attacks. The two significant contributions of this paper are as follows. The first contribution is incorporating a fuzzy unique transformation method within the architecture of a deep model to secure the deep model against the adversarial attacks. The second contribution is to provide a comprehensive study of the performance of the proposed model to classify the COVID-19 cases under the various adversarial attacks.

The organization of the paper is as follows: Section II presents the brief introduction of the adversarial attack and fuzzy set theory. Section III presents details of the methodology used to train the secure deep model using FUIT transformed images to prevent the adversarial attacks. Section IV presents experiments, results and ablation study, and finally, conclusions and future work are presented in Section V.

II. Preliminaries

This section presents a brief introduction of adversarial attack and fuzzy set.

A. Adversarial Attacks

The major aim of adversarial attacks [24] is to modify pixel values by small amount $\epsilon$. The changes that occur in a modified image are invisible for humans but well understood by deep learning models. If $f$ denotes a function that represents a deep model with parameters $\theta$ learned using input image $X$ and label $y$

$$y = f(X, \theta)$$  \hspace{1cm} (1)

After adding small perturbation to image pixels the adversarial input image $X'$ satisfies the following condition:

$$||X' - X|| \leq \epsilon$$  \hspace{1cm} (2)

$$y' = f(X', \theta)$$  \hspace{1cm} (3)

When the model is evaluated against adversarial image $X'$, then $y \neq y'$, that results in a degradation in the model’s performance. The phenomenon of a decrease in the classification rate of the model to classify the images is known as the adversarial attack that easily fools the model to misclassify the input image modified using a small perturbation $\epsilon$. In this paper, our primary aim is to prevent the deep model against adversarial attacks.
B. Fuzzy Set

A set whose every element have membership value is known as fuzzy set \(F\) \[42\],

\[
\tilde{F} = \{(x, \mu_F(x)) \mid x \in U\}
\] (4)

Where \(\tilde{F}\) is a fuzzy set with element \(x\) and membership value \(\mu\). Here \(\mu_F(x)\) denotes membership value of \(x\) with respect to fuzzy set \(\tilde{F}\). The value of \(\mu\) always lies in between 0 to 1. U is an universe of information.

III. METHODOLOGY

This section presents details of the Fuzzy Unique Image Transformation (FUIT) technique and methodology used to build a secure deep model against the adversarial attacks.

A. Fuzzy Unique Image Transformation (FUIT)

FUIT creates fuzzy sets from the given range of values of the image pixels. In an image where pixel values lie in a range \((U)\) from 0 to 255. We create the \(R\) fuzzy sets that use a triangular membership function \[42\] (as shown in Eq.(5)) to compute the membership value \((\mu)\) of the given pixel. The created fuzzy sets downsmaple the image pixels into an interval of range in between 1 to \(R\). The new transformed image has pixel values between 1 to \(R\). The FUIT technique performs discretization of the values of the image pixel into an interval \([1, R]\).

Algorithm 1 Fuzzy Unique Image Transformation (FUIT)

Require: Input Image \((X)\), \(R\)-Fuzzy Sets
1: Initialize \([\text{rows, cols}] = \text{size}(X)\)
2: for \((i=1) \& (i<=\text{rows})\) do
3: for \((j=1) \& (j<=\text{cols})\) do
4: for \((z=1) \& (z<=R)\) do
5: \(Mv[r]=\mu_x(X^v)\)
6: \(\text{Here, } Mv[1\sim R]\) = An array of membership values
7: end for
8: \([\mu_{max}, \text{Index}] = \text{Max}(Mv)\)
9: \(X^F = \text{Index}\)
10: end for
11: end for
12: Return \(X^F\) (Output FUIT Transformed Image)

Algorithm \([1]\) shows various steps of FUIT transformation of the image \(X\). In this paper a triangular membership function is used to compute membership values of a pixel for the given fuzzy set.

\[
\mu(x, p, q, r) = \begin{cases} 
0 & x \leq p \\
\frac{x-p}{q-p} & p \leq x \leq q \\
\frac{r-x}{r-q} & q \leq x \leq r \\
0 & r \leq x 
\end{cases}
\] (5)

Eq.(5) shows the triangular membership function with three parameters \(p, q, r\) and input \(x\). \(\mu\) denotes triangular membership value.

For an example, consider an image of size \((3*3)\) having values \([78, 61, 120, 236, 222, 40, 10, 11, 15]\) as shown in Fig.(1). The image is transformed into FUIT image using the Algorithm \([1]\). The new image created after the FUIT technique has pixel values \([4, 3, 6, 12, 11, 2, 1, 1, 1]\) based on membership values \([0.96, 0.72, 0.80, 0.88, 0.64, 0.67, 0.80, 0.88, 0.80]\) computed from the created fuzzy sets. Fig.(2) shows various fuzzy sets created for the FUIT transformation. After the previous image transformed into an adversarial image the new pixel values become \([81, 63, 123, 241, 222, 40, 17, 15, 17]\). After applying the FUIT algorithm the new transformed image becomes \([4, 3, 6, 12, 11, 2, 1, 1, 1]\) with membership values \([0.72, 0.56, 0.56, 0.72, 0.64, 0.64, 0.64, 0.64, 0.64]\).

It is clear from the Fig.(1) that the image persists its own characteristics in both the situations (for a clean image or under attack). The characteristics of the images can be expressed in terms of the number of unique pixel values \((V)\). For a clean image \(V = [4, 3, 6, 12, 11, 2, 1]\) and for an adversarial image \(V = [4, 3, 6, 12, 11, 2, 1]\). In both the cases \(||V|| = 7\). The FUIT transformation prevents the increase in the value of \(||V||\) due to adversarial attack. In other words, a variance of \(||V||\) remains the same for the clean image and image under adversarial attack. The increase in the variance of \(||V||\) due to adversarial attack easily fools the deep model and results in a high misclassification rate. The FUIT transformation is deployed before forwarding the image as input to the deep model. The model easily learns the FUIT transformed images and shows secure nature towards the adversarial attacks. All the training images are prepossessed with FUIT technique while training of the deep model and all test images i.e., clear images or adversarial images, are also pre-processed with FUIT technique before classification by the deep model. Fig.(3) and Fig.(4) show the overall flow of classification and defense mechanism used against the adversarial attacks on deep COVID-19 model using the proposed framework.

Fig. 1: Various steps of Fuzzy Unique Image Transformation
IV. EXPERIMENTS AND RESULTS

This section presents a description of datasets and the details of the model and results compared with the previous proposal for COVID-19.

A. COVID-19 Chest X-Ray Image Dataset

The performance of the proposed method is first evaluated on COVID-19 Chest X-Ray Image Dataset [43]. The data set includes a collection of chest X-Ray images of people belonging to Normal, Pneumonia, and COVID-19 classes. Several contributions from people belonging to different places
increase the size of the dataset. At the time of this study, the dataset contains a total of 1125 images. Among the available 1125 images, 500 images belong to Normal Class, 500 images belong to people suffering from pneumonia, and the remaining 125 images belong to people infected from COVID-19. The study followed 5-fold cross-validation to evaluate the performance of the proposed framework. Fig. 5 shows sample images of the chest X-Ray of persons belong to normal, pneumonia, and COVID-19 classes.

### B. Model Description

Total of nine models (M1-M9) are trained using weights initialized with three different pre-trained models i.e. Resnet-18 [44], VGG-16 [45] and GoogLeNet [46] respectively. Table I shows details of the nine models prepared for comparison. The models (M1-M9) are further evaluated against the adversarial images created using six different types of attacks exist in literature. Table II shows the six different attacks that are used in this study. Parameters of all the six attacks mentioned in the Table II are shown in Table III.

#### TABLE I: Description of Different Models Developed for Chest X-Ray Dataset

| Model | Dataset | Image Type | Pre-Trained Model |
|-------|---------|------------|-------------------|
| M1    | Chest X-Ray | Clean Image | Resnet-18 |
| M2    | Chest X-Ray | Clean Image | VGG-16 |
| M3    | Chest X-Ray | Clean Image | GoogLeNet |
| M4    | Chest X-Ray | FUIT Transformed Image | Resnet-18 |
| M5    | Chest X-Ray | FUIT Transformed Image | VGG-16 |
| M6    | Chest X-Ray | FUIT Transformed Image | GoogLeNet |
| M7    | Chest X-Ray | Discretization Transformed Image | Resnet-18 |
| M8    | Chest X-Ray | Discretization Transformed Image | VGG-16 |
| M9    | Chest X-Ray | Discretization Transformed Image | GoogLeNet |

#### TABLE II: Details of Different Adversarial Attacks

| S.No. | Attacks                  |
|-------|--------------------------|
| 1     | Deep Fool                |
| 2     | Fast Gradient Sign Attack (FGSM) |
| 3     | Basic Iterative Method (BIM) |
| 4     | Carlini & Wagner (CW)    |
| 5     | Projected Gradient Descent With Random Start (PGD-r) |
| 6     | Projected Gradient Descent Without Random Start (PGD) |

#### TABLE III: Parameters of Six Adversarial Attacks

| Attack          | Parameters                      |
|-----------------|---------------------------------|
| PGD             | $\epsilon=0.3, \alpha=4/255, \text{Steps}=40$ |
| PGD-r           | $\epsilon=0.3, \alpha=4/255, \text{Steps}=40, \text{Random Start}=True$ |
| FGSM            | $\epsilon=0.008$ |
| CW              | $C=2, Kappa=2, \text{Steps}=500, \text{learning rate}=0.01$ |
| Deep Fool       | $\epsilon=8/255, \alpha=1/255, \text{steps}=10$ |

#### C. Experimental Settings

The early stop technique is used to train the models, and the maximum epoch is set as 150. The learning rate is 0.001, and the batch size is selected as 32. All experiments are conducted on Ubuntu 16.04 LTS operating system with 16 GB RAM and NVIDIA GM107M 4 GB GPU. All scripts are developed using an open-source Pytorch 1.4 library. All images are resized to a size required by the three pre-trained models used to initialize the deep model weights. The deep model is trained using an Adam optimization [36].

#### D. Loss Function

A loss function $\mathcal{L}$ is selected as cross entropy loss as shown in Eq. (6). Here $x$ is an input and $C$ is class label, $k$ is the total number of classes.

$$
\mathcal{L}(x, C) = -\log \left( \frac{\exp (x|C|)}{\sum_k \exp (x|k|)} \right)
$$

#### E. Results on Chest X-Ray Dataset

As mentioned earlier for the comparative analysis total of nine models are developed. We evaluated the performance of the proposed model in two settings. Models are trained for binary and three class classification scenarios. In the case of the binary classification, images belong to COVID-19 and pneumonia classes are considered to come from the same class. Initially, M1, M2, and M3 are trained and tested on the clean chest X-ray images. Table IV shows the performance of the three models for binary and three class classification. For the binary classification, the model M1 yields the highest mean accuracy of 97.28%, and the model M3 shows the lowest mean accuracy of 96.84%. The model M2 achieves the accuracy of 97.14%. For the three-class classification, the model M1 shows the highest mean accuracy of 88.12%. The model M3 shows the lowest mean accuracy of 87.03%.

#### TABLE IV: Classification Accuracy of Different Models ON COVID-19 Chest X-ray Dataset

| Model | Accuracy (%) |
|-------|--------------|
| M1    | 97.28 ±0.21  |
| M2    | 97.14 ±0.30  |
| M3    | 96.84 ±0.33  |

#### TABLE V: Adversarial Attack on Models Trained Using Clean COVID-19 Chest X-Ray Images

| Attack          | Accuracy (Binary Class) | Accuracy (Three Class) |
|-----------------|-------------------------|------------------------|
| PGD             | 54.34 51.17 50.46       | 47.12 46.74 44.58      |
| PGD-r           | 50.12 48.31 48.12       | 40.17 39.51 39.23      |
| FGSM            | 47.12 46.18 46.05       | 39.14 38.74 37.12      |
| CW              | 10.74 10.25 10.05       | 9.81 9.19 8.89         |
| Deep Fool       | 14.61 13.92 13.84       | 12.29 11.81 11.01      |
| BIM             | 9.17 8.69 8.61          | 8.15 7.10 7.01         |

After evaluation of the performance of the models M1,M2 and M3 on the clean Chest X-Ray images, the models are tested against the adversarial images generated from six different attacks, as listed in the Table II. Table V shows the performance of the three models for binary and three class.
classification. The model M1 shows an accuracy of 54.34% in the presence of a PGD attack. The model M1 shows the accuracy of 50.12%, 47.12%, 10.74%, 14.61%, and 9.17% for PGD-r, FGSM, CW, Deep Fool, and BIM attacks, respectively. The BIM attack is the most successful attack that results in the lowest accuracy of 9.17%. For the same attack, M2 and M3 show an accuracy of 8.69% and 8.61%, respectively. For three-class classification, the lowest accuracy of M1, M2, and M3 are 8.15%, 7.10%, and 7.01%, respectively. It is clear from the Table [V] that the models trained on clean images perform poorly and highly insecure against the adversarial images generated by the six adversarial attacks.

Again three models M4, M5 and M6 are trained for binary and three class-classifications using the images obtained after the FUIT transformation. Table [VI] shows the accuracy of models M4, M5, and M6 when trained and test on FUIT transformed images. The accuracy of model M4 is 96.81% and 87.25% for binary and three class classification and little lower than the model M1. It is clear from Table [VI] that the FUIT transformed images are learnable by the deep model.

The models M4, M5, and M6 are tested against the adversarial images generated using the six attacks. It is worth noting here all the training, test, or adversarial images are first transformed into FUIT technique and then given as input

---

**TABLE VI:** Classification Accuracy of Different Models ON FUIT Transformed COVID-19 Chest X-ray Dataset

| Model (FUIT) | Accuracy (Binary Class) | Accuracy (Three Class) |
|-------------|--------------------------|-------------------------|
|             | M4          | M5          | M6          | M4          | M5          | M6          |
| Binary Class|             |             |             |
| PGD        | 96.54 ±0.14 | 96.20 ±0.26 | 96.03 ±0.31 |
| PGD-r      | 96.47 ±0.13 | 96.13 ±0.26 | 96.03 ±0.31 |
| FGSM       | 96.49 ±0.17 | 96.17 ±0.26 | 96.03 ±0.31 |
| CW         | 96.12 ±0.14 | 95.89 ±0.26 | 96.03 ±0.31 |
| Deep Fool  | 95.91 ±0.14 | 95.49 ±0.26 | 96.03 ±0.31 |
| BIM        | 95.31 ±0.14 | 95.28 ±0.26 | 96.03 ±0.31 |

---

**TABLE VII:** Adversarial Attack on Models Trained Using FUIT Transformed COVID-19 Chest X-Ray Images

| Attack  | Accuracy (Binary Class) | Accuracy (Three Class) |
|---------|--------------------------|-------------------------|
|         | M4          | M5          | M6          | M4          | M5          | M6          |
| PGD     | 96.54        | 96.20        | 95.97        | 87.23        | 86.81        | 85.49        |
| PGD-r   | 96.47        | 96.13        | 95.59        | 87.12        | 86.79        | 85.41        |
| FGSM    | 96.49        | 96.17        | 95.42        | 87.09        | 86.70        | 85.27        |
| CW      | 96.12        | 95.89        | 95.14        | 86.73        | 86.67        | 85.26        |
| Deep Fool | 95.91       | 95.49        | 95.06        | 86.62        | 86.60        | 85.19        |
| BIM     | 95.31        | 95.28        | 95.01        | 86.59        | 86.47        | 85.16        |
to the deep model. Table VII shows the accuracy achieved by the three models in the presence of six adversarial attacks. The model M4 shows the highest accuracy of 96.54% and 87.23% for the binary and three class classification when tested for the PDG attack. The same model also shows an accuracy of 95.13% and 86.59% against the most successful attack i.e., BIM attack. It is clear from Table VII that the proposed FUIT technique prevents the deep model against the adversarial attacks and persist high accuracy to classify the COVID-19 cases when attacked with the adversarial images. Table VIII and Table IX show a comparison of the developed models with the state-of-the-art (SOTA) methods for the diagnosis of COVID-19 cases using the chest X-Ray images. The accuracy of the proposed FUIT transformation-based model is comparable to SOTA models and provides reliable security against adversarial attacks. Fig.6 and Fig.7 show comparison of different models to detect the COVID-19 cases in binary and three class classification scenarios respectively.

TABLE VIII: Comparison of Accuracy of SOTA Methods for Binary Classification

| Model               | Mean (%) |
|---------------------|----------|
| M1                  | 97.28    |
| M2                  | 97.14    |
| M3                  | 96.84    |
| M4                  | 96.81    |
| M5                  | 96.27    |
| M6                  | 96.03    |
| Ozturk et al. [7]   | 98.08    |
| Khan et al. [8]     | 99.01    |
| Apostolopoulos et al. [11] | 98.75   |
| Wang et al. [9]     | 92.40    |
| Hemdan et al. [49]  | 90       |
| Narnin et al. [12]  | 97       |

TABLE IX: Comparison of Accuracy of SOTA Methods for Three Class Classification

| Model               | Mean (%) |
|---------------------|----------|
| M1                  | 88.21    |
| M2                  | 87.81    |
| M3                  | 87.03    |
| M4                  | 87.25    |
| M5                  | 86.95    |
| M6                  | 86.12    |
| Ozturk et al. [7]   | 87.02    |
| Khan et al. [8]     | 89.50    |
| Apostolopoulos et al. [11] | 92.85   |
| Wang et al. [9]     | 90.60    |

F. Ablation Study and Discussion

Apart from evaluating the performance of a deep model on the FUIT processed images, the study of the performance of the deep model trained on images transformed using the typical discretization is also evaluated. Table X shows the accuracy of models M7, M8, and M9 trained for binary and three class classification using the images transformed with the typical discretization. In typical discretization, the range of value of pixels in images is divided into intervals. For the standard discretization-based transformation, each time pixel value is divided by L and floor value is computed to know the interval to which the pixel value belongs. For an example, if value of the L is 32, then the total number of intervals is equal to 7 when pixels value has a range between 0 to 255. In this study, we set the value of L as 32. The normal discretization is a hard assignment of intervals, and the proposed FUIT technique is a soft assignment of intervals. The model M7 shows the highest mean accuracy of 96.25% and 86.91% for the binary and three class classification, respectively. The models M7, M8, and M9 show less accuracy than the models M4, M5, and M6. The soft assignment of intervals using the FUIT technique is capable of dealing with

Fig. 6: Comparison of Accuracy of Different Models for Binary Class Classification to Classify Chest X-Ray Dataset

Fig. 7: Comparison of Accuracy of Different Models for Three Class Classification to Classify Chest X-Ray Dataset
TABLE XI: Adversarial Attack on Models Trained Using discretization-Based Transformed COVID-19 Chest X-Ray Images

| Attack | Accuracy (Binary Class) | Accuracy (Three Class) |
|--------|-------------------------|------------------------|
|        | M7          | M8          | M9          | M7          | M8          | M9          |
| PGD    | 96.10       | 95.97       | 95.51       | 86.53       | 86.12       | 85.41       |
| PGD-r  | 96.07       | 95.91       | 95.49       | 86.51       | 86.10       | 85.39       |
| FGSM   | 95.86       | 95.80       | 95.31       | 86.12       | 86.09       | 85.21       |
| CW     | 95.81       | 95.79       | 95.29       | 86.10       | 85.91       | 85.17       |
| Deep Fool | 95.77   | 95.61       | 95.18       | 85.97       | 85.89       | 85.12       |
| BIM    | 95.70       | 95.47       | 94.97       | 85.91       | 85.81       | 84.96       |

uncertainty occurs during the assignment of the pixels into the intervals, thus resulting in higher accuracy of the model M4 compared to the model M7. Table [XI] shows the accuracy of the models M7, M8, and M9 for binary and three class classification in the presence of the six adversarial attacks. The models trained on images transformed using the typical discretization approach also show high accuracy to prevent the deep COVID-19 model against the adversarial attacks but show less accuracy as compared to models trained using the FUIT transformed images. It is clear from the Table [VII] and Table [XI] that the models trained using the FUIT transformed images are more secure against the adversarial attacks while diagnosis of the COVID-19 cases. All the nine models show high classification accuracy for the binary class classification and less accuracy for three class-classifications. The COVID-19 cases share similar symptoms with the pneumonia cases, which results in less classification accuracy in the case of the three-class classification.

G. Results on CT Image Dataset

The proposed model is also evaluated on second available CT Scan Image Dataset [50] for the diagnosis of COVID-19. The dataset contains 398 images for normal patients and 399 images for the patients suffer from COVID-19. Total of nine models (M10-M18) are trained by initialize the weighted using the three pre-trained models Resnet-18 [44], VGG-16 [45] and GoogLeNet [46]. Table XII shows a description of developed models to evaluate the performance of the proposed method on CT Image Dataset. Classification of COVID-19 cases in CT images is a problem of binary classification. The evaluation is performed by applying 5-fold cross-validation technique. Table XIII shows the accuracy of nine models developed to classify the COVID-19 cases. Fig. [8] shows sample images belong to normal and COVID-19 cases in the CT image dataset.

The model M10 achieves the highest mean accuracy of 89.19%. The model M18 shows the lowest mean accuracy of 87.77%. The developed models are tested to classify the COVID-19 cases in the presence of six adversarial attacks from the Table [XI]. The value of L is set as 32 for models M16, M17, and M18. Table XIV shows the accuracy of models to classify the COVID-19 cases under the six attacks. The models M10, M11, and M12 show degradation in classification accuracy when tested against the adversarial CT images. The BIM attack is again the most successful attack that drops the accuracy of model M10 from 89.19% to 8.12%. The drop in classification accuracy for model M11 and M12 is 81.07% and 79.86%, respectively. The highest classification accuracy of models M10, M11, and M12 under PGD attack are 48.19%, 47.81%, and 46.65%, respectively. However, the models M13, M14, and M15 show the accuracy of 86.12% and 86.01% and 85.97% respectively under the PGD attack.

TABLE XII: Description of Models Developed for CT Image Dataset

| Model | Dataset | Image Type | Pre-Trained Model |
|-------|---------|------------|-------------------|
| M10   | CT Image | Clean Image | Resnet-18 |
| M11   | CT Image | Clean Image | VGG-16 |
| M12   | CT Image | Clean Image | GoogLeNet |
| M13   | CT Image | FUIT Transformed Image | Resnet-18 |
| M14   | CT Image | FUIT Transformed Image | VGG-16 |
| M15   | CT Image | FUIT Transformed Image | GoogLeNet |
| M16   | CT Image | Discretization Transformed Image | Resnet-18 |
| M17   | CT Image | Discretization Transformed Image | VGG-16 |
| M18   | CT Image | Discretization Transformed Image | GoogLeNet |

TABLE XIII: Performance of Different Models to Classify Images of CT Image Dataset

| Model | Accuracy (%) |
|-------|--------------|
| M10   | 89.19 ±0.13  |
| M11   | 89.12 ±0.18  |
| M12   | 88.37 ±0.22  |
| M13   | 88.31 ±0.17  |
| M14   | 88.25 ±0.24  |
| M15   | 88.07 ±0.19  |
| M16   | 88.03 ±0.12  |
| M17   | 87.98 ±0.11  |
| M18   | 87.77 ±0.21  |

TABLE XIV: Performance of Different Models Under Six Different Adversarial Attacks

| Attack | Accuracy (%) |
|--------|--------------|
|        | Models       |
|        | M10 | M11 | M12 | M13 | M14 | M15 | M16 | M17 | M18 |
| PGD    | 48.19 | 47.81 | 46.65 | 86.12 | 86.01 | 85.97 | 86.02 | 85.58 | 85.37 |
| PGD-r  | 47.52 | 47.39 | 46.01 | 86.07 | 86.01 | 85.98 | 85.87 | 85.81 | 85.62 |
| FGSM   | 43.71 | 43.59 | 41.92 | 86.01 | 86.01 | 85.93 | 85.81 | 85.77 | 85.61 |
| CW     | 8.89  | 9.71  | 9.99  | 85.97 | 85.87 | 85.91 | 85.80 | 85.72 | 85.57 |
| Deep Fool | 12.19 | 12.01 | 11.86 | 85.92 | 85.83 | 85.78 | 85.71 | 85.70 | 85.56 |
| BIM    | 8.12  | 8.05  | 7.91  | 85.21 | 85.19 | 85.01 | 85.02 | 85.01 | 85.01 |

It is clear from the Table XIV that the models M13, M14 and M15 are more secure towards the six adversarial attacks and maintain the high accuracy to classify the COVID-19 cases. The accuracy of the models M13, M14 and M15 for the BIM attack are 85.21%, 85.19% and 85.01% respectively. However the accuracy of the models M16, M17 and M18 are 85.02%, 85.01% and 85.01% respectively. The accuracy of the models M16, M17 and M18 are little less than the model M13, M14 and M15 which shows the FUIT technique efficiently deal with uncertainty created during the downsampling of the image pixels into an interval. The performance of the models
Fig. 8: Sample Images of Normal, COVID-19 Cases From CT Image Dataset [43]

Fig. 9: Clean and Adversarial CT Image of COVID-19 Class Misclassified By Model M10 Under PGD Attack, (a) Clean COVID-19 CT Image Classified by Model M10 with \( \text{Prob}(\text{COVID}) = 0.85 \), \( \text{Prob}(\text{Normal}) = 0.15 \), (b) Adversarial COVID-19 CT Image Misclassified with \( \text{Prob}(\text{COVID}) = 0.41 \), \( \text{Prob}(\text{Normal}) = 0.59 \)

M13, M14 and M15 as shown in the Table [XIV] verifies that FUIT transformed images make the deep model more secure against the adversarial attacks and helpful to develop more reliable deep models for the diagnosis of COVID-19 cases.

Fig. (9a) shows a clean COVID-19 CT image correctly classified as COVID-19 case with a class probability of 0.85 by the model M10. On the other hand when the model is attacked with PGD attack the model classifies the same image as a Normal case with a class probability of 0.59. Fig. (9b) shows the adversarial image of the COVID-19 case (adversarial image of the image shown in Fig. (9a)). In presence of PGD attack the class probability of COVID-19 decreases to 0.41. The difference between the clean COVID-19 CT image and adversarial COVID-19 CT is visually unrecognizable by humans but well recognized by the deep model. Fig. (9) shows comparison of these two images when classified by model M10 in presence of the PGD attack. Table [XV] shows comparison of the proposed model with SOTA methods to classify the COVID-19 cases using the CT images. Fig. (10) shows comparison of accuracy of different models to classify the COVID-19 cases in the CT image dataset.

V. CONCLUSION

In this paper, we presented a novel fuzzy unique image transformation (FUIT) technique as a pre-processing step that prevents the COVID-19 deep model against the adversarial attacks. The FUIT technique downsamples the image pixels into an interval by using the created fuzzy sets. The FUIT technique prevents an increase in the variance of the number of unique pixels of the given image. This results in an equal number of unique pixels values in the clean and adversarial images. The deep model trained using the FUIT transformed images shows robust and secure performance against the adversarial attacks. The experiment and results on two available COVID-
TABLE XV: Comparison of Proposed Model With SOTA Methods to Classify CT Images

| Model       | Accuracy (%) |
|-------------|--------------|
| M10         | 89.19        |
| M11         | 89.12        |
| M12         | 88.37        |
| M13         | 88.31        |
| M14         | 88.25        |
| M15         | 88.07        |
| M16         | 88.03        |
| M17         | 87.98        |
| M18         | 87.77        |
| Bernheim et al. [6] | 88 |
| Angelov et al. [51] | 88.60 |

![Fig. 10: Comparison of Accuracy of Different Models for Binary Class Classification to Classify CT Image Dataset](image)

19 diagnosis datasets validate that the model maintains high accuracy to classify the COVID-19 cases in various non-targeted adversarial attacks. Moreover, the proposed model trains more secure deep models that prevent the COVID-19 deep model from the adversarial attacks.

The study is performed using datasets with significantly fewer images, which could be one drawback of this study. In future, the models will be trained on more images collected from other publicly available repositories and nearby local hospitals. Besides, an inspection of the FUIT to develop the deep models to classify the images received from various research domains can be considered a natural extension of this study.

REFERENCES

[1] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei et al., “A new coronavirus associated with human respiratory disease in china,” Nature, vol. 579, no. 7798, pp. 265–269, 2020.

[2] C. Collaborative, “Global guidance for surgical care during the covid-19 pandemic,” The British journal of surgery, 2020.

[3] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, “Correlation of chest ct and rt-PCR testing in coronavirus disease 2019 (COVID-19) in china: a report of 1014 cases,” Radiology, p. 200642, 2020.

[4] J. Shuja, E. Alanazi, W. Alasmary, and A. Alashaikh, “COVID-19 open source data sets: A comprehensive survey,” medRxiv, 2020.

[5] Z. Y. Zu, M. D. Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. M. Lu, and L. J. Zhang, “Coronavirus disease 2019 (COVID-19): a perspective from china,” Radiology, p. 200490, 2020.

[6] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z. A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li et al., “Chest ct findings in coronavirus disease-19 (COVID-19): relationship to duration of infection,” Radiology, p. 200463, 2020.

[7] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, “Automated detection of covid-19 cases using deep neural networks with x-ray images,” Computers in Biology and Medicine, p. 103792, 2020.

[8] A. I. Khan, J. L. Shah, and M. M. Bhat, “Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images,” Computer Methods and Programs in Biomedicine, p. 105581, 2020.

[9] L. Wang and A. Wong, “Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images,” arXiv, pp. arXiv–2003, 2020.

[10] A. Shoiebi, M. Khodatars, R. Alizadehsani, N. Ghassemi, M. Jafari, P. Moridian, A. Khadem, D. Sadeghi, S. Hussain, A. Zare et al., “Automated detection and forecasting of covid-19 using deep learning techniques: A review,” arXiv preprint arXiv:2007.10785, 2020.

[11] I. D. Apostolopoulos and T. A. Mpesiana, “COVID-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks,” Physical and Engineering Sciences in Medicine, p. 1, 2020.

[12] A. Narin, C. Kaya, and Z. Pamuk, “Automatic detection of coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks,” arXiv preprint arXiv:2003.10849, 2020.

[13] S. Misra, S. Jeon, S. Lee, R. Managuli, I.-S. Jang, and C. Kim, “Multi-channel transfer learning of chest x-ray images for screening of covid-19,” Electronics, vol. 9, no. 9, p. 1388, 2020.

[14] Z. Han, B. Wei, Y. Hong, T. Li, J. Cong, X. Zhu, H. Wei, and W. Zhang, “Accurate screening of covid-19 using attention based deep 3d multiple instance learning,” IEEE Transactions on Medical Imaging, 2020.

[15] J. Wang, Y. Yao, Y. Wen, H. Lu, H. Luo, Y. Xiang, X. Li, C. Liu, and D. Qian, “Prior-attention residual learning for more discriminative covid-19 screening in ct images,” IEEE Transactions on Medical Imaging, 2020.

[16] V. Sharma and C. Dyreson, “COVID-19 detection using residual attention network an artificial intelligence approach,” arXiv preprint arXiv:2006.16106, 2020.

[17] A. Mangal, S. Kalia, H. Rajgopal, K. Rangarajan, V. Namboodiri, S. Banerjee, and C. Arora, “Covidaid: COVID-19 detection using chest x-ray,” arXiv preprint arXiv:2004.09803, 2020.

[18] A. Abbas, M. M. Abdelsamea, and M. Gaber, “4s-dt: Self supervised super sample decomposition for transfer learning with application to covid-19 detection,” arXiv preprint arXiv:2007.11450, 2020.

[19] Y. Li, W. Dong, J. Chen, S. Cao, H. Zhou, Y. Zhu, J. Wu, L. Lan, W. Sun, T. Qian et al., “Efficient and effective training of covid-19 classification networks with self-supervised dual-track learning to rank,” IEEE Journal of Biomedical and Health Informatics, 2020.

[20] M. Karim, T. Döhmen, Rebholz-Schuhmann, S. Decker, M. Cochez, and O. Beyan, “Deepcovidexplainer: Explainable covid-19 predictions based on chest x-ray images,” arXiv preprint arXiv:2004.04582, 2020.

[21] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, “Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays,” Computer Methods and Programs in Biomedicine, p. 105608, 2020.

[22] R. K. Singh, R. Pandey, and R. N. Babu, “Covidscreen: Explainable deep learning framework for differential diagnosis of covid-19 using chest x-rays,” 2020.

[23] Y.-H. Wu, S.-H. Gao, J. Mei, J. Xu, D.-P. Fan, C.-W. Zhao, and M.-M. Cheng, “Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation,” arXiv preprint arXiv:2004.07054, 2020.

[24] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and P. Robi, “Evasion attacks against machine learning at test time,” in Joint European conference on machine learning and knowledge discovery in databases. Springer, 2013, pp. 387–402.

[25] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, “Scalable object detection using deep neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 2147–2154.

[26] N. Mehdiyev, J. Lahann, A. Emrich, D. Enke, P. Fettke, and P. Loos, “Time series classification using deep learning for process planning: a case from the process industry,” Procedia Computer Science, vol. 114, pp. 242–249, 2017.
[27] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, “A survey of deep neural network architectures and their applications,” *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[28] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, “Deep learning for visual understanding: A review,” *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[29] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.

[30] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[31] H. Hirano, K. Koga, and K. Takemoto, “Vulnerability of deep neural networks for detecting covid-19 cases from chest x-ray images to universal adversarial attacks,” *arXiv preprint arXiv:2005.11061*, 2020.

[32] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, “Adversarial attacks on medical machine learning,” *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[33] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[34] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[35] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[36] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[37] V. Sharma, A. Kalra, S. C. Vaibhav, L. Patel, and L.-P. Morency, “Attend and attack: Attention guided adversarial attacks on visual question answering models,” in *Proc. Conf. Neural Inf. Process. Syst. Workshop Secur. Mach. Learn*, 2018.

[38] D. J. Miller, Z. Xiang, and G. Kesidis, “Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks,” *Proceedings of the IEEE*, vol. 108, no. 3, pp. 402–433, 2020.

[39] L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning,” *arXiv preprint arXiv:1712.04621*, 2017.

[40] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, “Countering adversarial images using input transformations,” *arXiv preprint arXiv:1711.00117*, 2017.

[41] P. Zhao, Z. Fu, Q. Hu, J. Wang et al., “Detecting adversarial examples via key-based network,” *arXiv preprint arXiv:1806.00580*, 2018.

[42] L. A. Zadeh, “Fuzzy sets,” *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[43] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, “Covid-19 image data collection: Prospective predictions are the future,” *arXiv preprint arXiv:2006.11988*, 2020.

[44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[45] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[47] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” *arXiv preprint arXiv:1607.02533*, 2016.

[48] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.

[49] E. E.-D. Hemdan, M. A. Shouman, and M. E. Karar, “Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images,” *arXiv preprint arXiv:2003.11055*, 2020.

[50] J. Zhao, Z. Zhang, X. He, and P. Xie, “Covid-ct-dataset: a ct scan dataset about covid-19,” *arXiv preprint arXiv:2003.13865*, 2020.

[51] P. Angelov and E. Almeida Soares, “Explainable-by-design approach for covid-19 classification via ct-scan,” *medRxiv*, 2020.