An Adversarial Attacks Resistance-based Approach to Emotion Recognition from Images using Facial Landmarks

Harisu Abdullahi Shehu[1], Will Browne [2] and Hedwig Eisenbarth[3]

***Abstract*— Emotion recognition has become an increasingly important area of research due to the increasing number of CCTV cameras in the past few years. Deep network-based methods have made impressive progress in performing emotion recognition-based tasks, achieving high performance on many datasets and their related competitions such as the ImageNet challenge. However, deep networks are vulnerable to adversarial attacks. Due to their homogeneous representation of knowledge across all images, a small change to the input image made by an adversary might result in a large decrease in the accuracy of the algorithm. By detecting heterogeneous facial landmarks using the machine learning library Dlib we hypothesize we can build robustness to adversarial attacks. The residual neural network (ResNet) model has been used as an example of a deep learning model. While the accuracy achieved by ResNet showed a decrease of up to 22%, our proposed approach has shown strong resistance to an attack and showed only a little ($< 0.3\%$) or no decrease when the attack is launched on the data. Furthermore, the proposed approach has shown considerably less execution time compared to the ResNet model.**

I. INTRODUCTION

Emotion recognition is an important area of research that has gained attention in the last two decades [1]. Information like nonverbal signals in human to human communication can be transmitted by facial expression [2], [3]. For instance, information can be conveyed via a scowl or a smile. It is considered as a natural way of trying to understand the psychological state of people during communication [4].

Nowadays, hospitals and healthcare systems are highly motivated to introduce autonomous robotic systems into highly critical healthcare operations. As such, these robots need to understand human emotions as well as their expectations in these close-proximity situations, so as to be able to work collaboratively in a socially-intuitive way [5].

Disciplines such as sociology [6], neuroscience [7], computer science [8], and psychology [9] have been producing research on facial expressions on different kinds of benchmark emotion images [10], [11], [12], [13]. Modern methods use deep learning (DL) algorithms as they have been widely used in intelligent systems such as home automation, robots, and autonomous vehicles, etc. to recognize human facial expressions [14].

*This work was not supported by any organization

[1]Harisu Abdullahi Shehu is with School of Engineering and Computer Science, Victoria University of Wellington, 6012 Wellington, New Zealand harisushehu@ecs.vuw.ac.nz

[2] Will Browne is with School of Engineering and Computer Science, Victoria University of Well ington, 6012 Wellington, New Zealand will.browne@vuw.ac.nz

[3]Hedwig Eisenbarth is with School of Psychology, Victoria University of Wellington, 6012 Wellington, New Zealand hedwig.eisenbarth@vuw.ac.nz

However, due to their consideration of the color distribution in pixels within a dataset, these algorithms are doing more of an image classification rather than an emotion recognition task. Also, because of their homogeneous representation of knowledge, they are vulnerable to a method called adversarial attack [15] where a change is made to the input with the intention of misleading the deep model to misclassify the input object [16]. Since DL models heavily rely on training data, a small change in the input or training data that is dissimilar to the train or test data might result in a large decrease in the overall accuracy of the algorithm. This is a problem that needs to be resolved as individual differences such as wearing of glasses, skin color et cetera should not be taken into consideration when analyzing emotion.

On the other hand, we know from the work of Ekman's facial action coding systems (FACS) [17] as well as Boyko et al.'s work [18] that emotion recognition is much more landmark-based. Therefore, using facial landmarks we would expect to get a much higher accuracy or at least equivalent accuracy even when a small change is made to the input image as the use of landmarks is more of a heterogeneous approach compared to the former which is a homogeneous level approach.

In this research, we aim to develop an adversarial resistance-based method to analyze emotion within faces in an image using landmarks. As the landmarks are extracted based on the facial patterns expressed in the image, the method should be able to generalize across the dataset due to its holistic representation of information.

We will test this using an adversarial attack and compare the method with the traditional method of analyzing emotion by deep learning models in which images are directly fed to the models giving them all the responsibility to do end-to-end learning. We chose the CK+ dataset as it contains video-frame images and used the last-half of frames of each sequence so as to have more data to be used for the experiment and apply both approaches on the database.

The rest of the paper is organized as follows: Section 2 explains the properties of the CK+ database, its emotion categories, and how it is used in this research. Section 3 explains how people use deep models to analyze emotion on different emotion databases and how certain research uses adversarial images to analyze emotion. In Section 4, we explain possible problems of feeding the classifier directly with extracted landmarks of images and how landmarks of faces are processed after they are extracted from images to avoid those problems. In addition, we explain the types of adversarial attacks used and why they are used in this re-

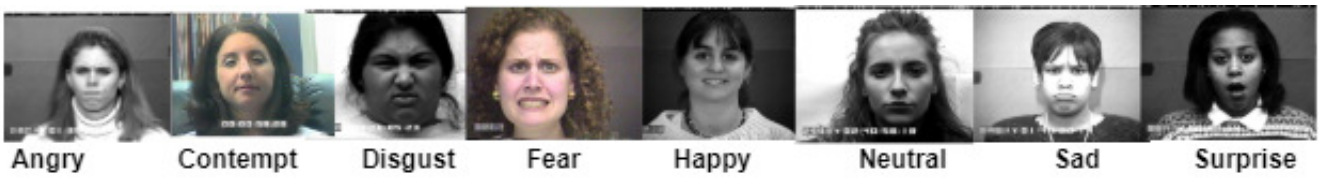

Fig. 1. Sample images of the CK+ database. Note: Certain images of the CK+ database are gray images. Therefore, all were converted to gray images

search. Section 5 presents the obtained results and compares the proposed approach to a deep learning method both before and after the adversarial attack has been launched on the data. In Section 6, we further discuss the obtained result and explain why training the deep learning model directly with images is not recommended in emotion recognition tasks. In Section 7, we conclude the paper and hint at further study.

II. Dataset

The extended Cohn-Kanade database (CK+) [19] is an extended version of the Cohn-Kanade [20] database of mainly posed facial expressions. Seven peak expressions; six basic (anger, disgust, fear, happy, sad, surprise) defined by Ekman [21] as well as contempt expressions were posed by 201 adults between the ages of 18 to 50 years. A total number of 593 sequences varying from 10 to 60 frames starting from neutral to peak expressions were captured from 123 subjects. Much research carried out in the field of emotion recognition was carried out on the six basic and neutral, which is the starting point of these dynamic expressions [22], [23].

In this research, a total of 3,368 images, which consists of the last-half of the frames of each sequence of the six basic expressions are used as the peak expressions and the first-two frames of each sequence are used as the neutral expression. Figure 1 shows sample images of the CK+ database.

III. Literature Review

Melaugh et al. [24] proposed an approach to classify emotion on certain portions of the face. Eight different regions; the eyes, mouth, right side of the face, left side of the face, right eye, left eye, the right side of the mouth and the left side of the mouth were cropped and tested using a single layer Convolution Neural Network (CNN) and the result achieved from the eight regions was compared with the result achieved using the full face region. In an attempt to test their methodology, all 213 images of the Japanese Female Facial Expressions (JAFFE) [25] database and 980 front face images of the KDEF [26] were used. As described in Section 2, all sequences of the CK+ database start from neutral to apex expressions where the expressions are expressed at peak. Melaugh et al. also used 469 images from the CK+ database, considering the last three frames of each sequence as the apex and the first frame of each sequence as the neutral expressions respectively. An accuracy of up to 89.4%, 87.32%, and 76.56% on KDEF, CK+, and JAFFE databases has been achieved respectively, with the full face region outperforming other extracted regions of the face. However, this research only used the peak frames of CK+ and the frontal face image of the KDEF database, so did not take into account other frames of the CK+ where the emotion is not expressed at peak and the side view images of the KDEF database. Thus, the produced model might perform badly on frames where the emotion is not expressed at peak and side-view images of the KDEF database.

A model to recognize facial expression based on transfer features from deep convolution networks (ConvNet) has been proposed [27]. High-level features from a trained deep ConvNet on a celebrity facial database (MSRA-CFW) [28] with 1580 face identification classes have been extracted. The first and last images of CK+ database, all JAFFE images, and frontal face images from KDEF and Pain expressions set from Psychological Image Collection at Stirling [11] were selected resulting in a total of 2062 images of seven emotion states (6 basic + neutral) and were used in the experiment, which led to an accuracy of 81.5% when the transfer feature is used together with the support vector machines (SVM) classifier. This approach used static as opposed to dynamic frames of the CK+ and used an artificial method of face selection to select all non-detected faces over automated selection methods. While the artificial method of face selection ensures that no face is omitted, it might be slow compared to an automated method, which will increase the overall turnaround time for the method to execute.

Tian et al. [29] proposed an approach based on a Secondary Information aware Facial Expression Network (SIFE-Net) to explore components without auxiliary labeling and a dynamic weighing strategy to teach the SIFE-Net. The method extracts secondary information from the image and trains the network with both expression labels and knowledge from the extracted information rather than training the images directly with one-hot encoded labels. Results show that the proposed SIFE-Net achieved state-of-the-art performance when tested on both CK+ and the Real-world Affective Faces (RAF-DB) [30] databases. Unlike the traditional algorithms, the proposed SIFE-Net not only learns from one-hot encoded labels but at the same time learns from the extracted information for sub-category knowledge. This is in addition to the knowledge already learned by the SIFE-Net method and might increase the overall accuracy of the method.

Le et al. [31] proposed an adversarial network-based method to perform facial expression recognition on occluded images. The method consists of a generator which complements occlusion in the generated images under three different

Fig. 2. Sample images of the CK+ database after landmarks detection. Note: Each green circle on the face denotes *(x, y)* coordinates of a particular landmark.

loss constraints, and an optimized discriminator that can distinguish between real and fake generated images by constructing an adversarial loss. Experimental analysis performed on RAF-DB with and without de-occlusion processing to verify the effectiveness of their method shows that the method outperforms the existing state-of-the-art methods. However, this method does not fully analyze the relationship between real occluded and non-occluded areas of the face. In addition to this, the generated method does not effectively recognize occluded facial images when more than 40% of the whole image is occluded.

Deep learning models [24], [27], [29] have been used to analyze emotion on images. However, feeding a deep learning model directly with images to perform emotion recognition tasks analyzes the color distribution rather than the pattern of the images, which may not generalize. This is anticipated to make the technique less resistant to an adversarial attack. In this research, we are going to address the issue of training deep models directly with images to analyze emotion by using landmarks within faces in images of the CK+ database to develop an adversarial attack resistant based approach.

IV. Methods

A. Hardware specification

A 9GB Graphical Processing Unit (GPU) device GeForce GTX 1080ti with CUDA version 10.2 is used in this research.

B. Method

- Pre-processing: Image pixel values were converted to an array and emotion labels were converted to integers. As certain images of the CK+ are grayscale images, all images were converted to grayscale images to ensure homogeneity of all images.
- Feature Extraction: Initially, we used Dlib library [32] which is an open-source machine learning software written in C++ to extract facial landmarks. Figure 2 shows an example of landmarks detected on selected images of the CK+ database. Feeding the classifier with facial landmarks directly extracted from an image might not give us the most accurate result we aim to obtain as certain faces of participants in the dataset might be located at a different location in the image. An example of this is shown in Figure 3 where the images are not co-located.
However, since the relationship between the *(x, y)* coordinates in each emotion expressed is expected to be the same, we find the average of each point (*x* and *y*) which we refer to as the central point and get the distance of each point relative to the central point as in Figure 4.

As shown in Figure 4, point C denotes the (*x* and *y*) coordinates of the central (average) point of the landmarks, point RB denotes the distance between the central point to a particular landmark point on the right eyebrow, point RE denotes the distance between the central point to a particular landmark point on the right eye, point LB denotes the distance between the central point to a particular landmark point on the left eyebrow, point LE denotes the distance between the central point to a particular landmark point on the left eye, point I denotes the distance between the central point to a particular landmark point in the inner mouth, point N denotes the distance between the central point to a particular landmark point on the nose, point M denotes the distance between the central point to a particular landmark point on the mouth and point J denotes the distance between the central point to a particular landmark point on the jaw.

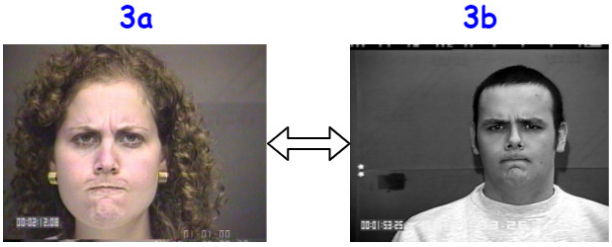

Fig. 3. Sample emotion expressed at different locations of an image. Note: While both expressions in the images denoted by 3a and 3b are the same (anger expressions). The face of the participant in 3a is shifted to the left side whereas the face of the participant in 3b is shifted to the right side of the image.

C. Adversarial Attack

An adversarial attack is a type of attack that is launched on an input by adding certain noise on it so as to fool the deep model from recognizing the input correctly. In our case, the input here is an image. An adversarial attack can be targeted and non-targeted. A targeted attack is when the deep model is fooled such that it classifies images of one class to a specific class or not to correctly classify an image of a particular class (e.g. classify anger images as fear images or classify anger images to any other class except for anger class). Conversely, a non-targeted adversarial attack aims to mislead the deep model to misclassify instances (images) and to reduce its accuracy but do not specify which class the deep model misclassifies the images to.
Furthermore, an adversarial attack can be performed based on different sets of knowledge. The white-box approach is where the target model is known to the adversary and the black-box is where the target model is not known. In this

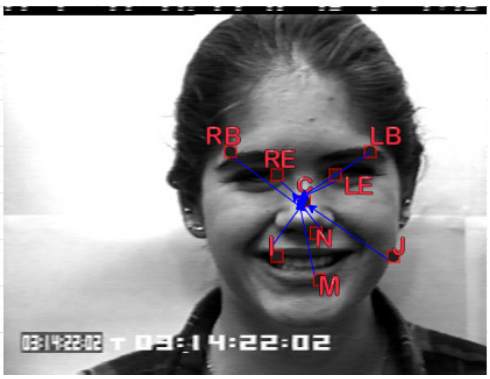

Fig. 4. Sample distance of landmark coordinates from the central point. Note: C here means central, RB means right eyebrow, RE means right eye, LE means left eye, LB means left eyebrow, N means nose, I means inner mouth, M means mouth, and J means Jaw.

research, we launch three different kinds of attacks on the images of the CK+ database. The fast gradient sign method attack [33], which is a white-box attack, and two custom-created black-box attacks that we refer to as *type A* and *type B* attack as shown in Figure 5.

- *Type A*: A deep model trained to recognize people's emotions should be able to recognize the emotion of any individual whether the person is wearing glasses or not. Due to the unavailability of an emotion database of people wearing glasses and the cost of producing real-world data, the *Type A* attack is performed by adding a rectangular bounding box on the left and right eyes of the original image instead of having glasses on the face as shown in Figure 5. Here we use a Type A attack as a targeted adversarial attack to train the images of a particular class and test the classifier with normal images. The aim is to mislead the classifier to classify the images of the class trained with *Type A* adversarial images as other classes except for the target class.
- *Type B*: This attack is affected by adding a rectangular bounding box on the jaw of each of the original images. A *Type B* attack is applied as a non-targeted attack to all test images to mislead the deep model to misclassify the data so as to reduce the overall accuracy of the model.

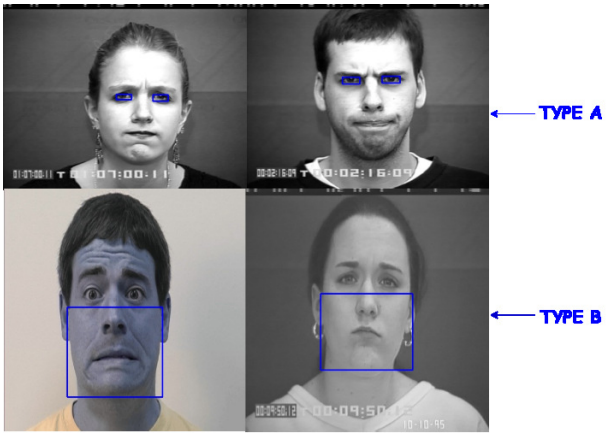

Fig. 5. Sample attack launch on images of the CK+ database. Type A attack is applied to the eyes whereas Type B attack is applied to the jaw.

- *Fast Gradient Sign Method (FGSM)*: This method creates an adversarial image using the gradient information of the neural network. It uses the gradients of the loss with respect to the input image to create a new image called the adversarial image that maximizes the loss. Equation (1) summarizes how FGSM is used here [34].

  Figure 6 shows an example of how FGSM is applied to the database. The first row shows an example of a *surprise* image predicted by the ResNet model. Initially, the model predicted the image as a *surprise* image with the confidence of 99.999% before the application of the FGSM method. However, while the picture remains the same class to a human after the application of the FGSM method with an ϵ (see equation 1) of 0.01, the unnoticeable change is enough to fool the ResNet model to predict the image as *anger* with a confidence level of 97.887%.
  Conversely, the *sad* image in the second row is predicted with a confidence of 100%. As such, the model loss is zero and as a result, the FGSM method could not disrupt the ResNet model and the resulting adversarial image remains the same as the original image, which in turn is predicted again as a *sad* expression with 100% confidence.

$$Adversarial_x = x + \epsilon * sign(\nabla_x J(\theta, x, y)) \quad (1)$$

where

  - $Adversarial_x$: Adversarial Image
  - x : Original Image
  - ϵ : Multiplier to ensure perturbations are small
  - J : Loss
  - θ : Model parameter
  - y : Original label

  The gradients of the loss are taken with respect to the input image, not the model parameter by finding

how much each pixel in the image contributes to the loss value. The model parameter remains constant as the model is no longer being trained. Hence, fooling an already trained model is the only goal here.

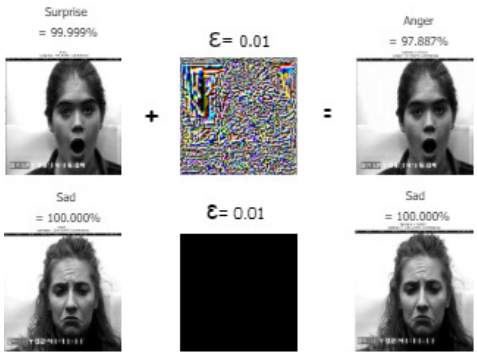

Fig. 6. Fast gradient sign method applied to test images. No changes are made to any image predicted with a confidence level of 100%

D. Classifiers

A supervised learning algorithm, Random Forest (RF), and a deep learning (DL) classifier, the residual neural network (ResNet), are the classifiers used in this research.

ResNet is chosen based on its performance obtained in the research of [35] and also because it adds zero new parameters; that is to say, we train exactly the same number of parameters as we would have had there been no residual connections. This is good because more parameters can lead to overfitting and fewer parameters lead to less training time. Also, as stated in the original paper, a ResNet with 34 layers only requires 18% of operations as a Visual Geometry Group (VGG) [36] with 19 layers (around half the layers of the ResNet) will require.

Similarly, RF is used based on its performance obtained in the research of [37] and also because they are ensemble methods that are averaged over many trees.

The Random Forest algorithm [38] constructs and trains multiple decision trees (DT) [39]. The algorithm improves its accuracy by collecting multiple decisions from the DTs it constructs. In general, the more trees used in RF, the better the result. However, at a certain point, since the expected variance decreases as the square root of the sample size, the cost of collecting more trees becomes higher than the benefit in accuracy obtained. For that reason, different choices for the number of trees such as 10, 30, 100, and 1000 were taken into consideration of which 100 was found to give the best accuracy result. As such, in this research, one hundred DTs are used to construct the RF algorithm.

A ResNet identity block is used to train the network from scratch. The description of the architecture is as follows: The number of parameters is approximately 281,000. While the number of filters is doubled at each stage, downsampling is performed on the feature map at each convolution (Conv) layer to reduce the size of the feature map to half the size of the original image in the first layer or half the size of whatever the feature map is in the previous Conv layer. The learning rate is scheduled to be reduced after each 80, 120, 160, and 180 epochs to avoid overfitting. Batch normalization normalizes each of the batches and a rectified linear unit (ReLU) is used as an activation function. Average pooling is applied after the addition of the original input to the feature map and reapplication of the activation function. Finally, the image is passed to a flatten and dense layer before Softmax is used to predict the class of the image. This set-up is exactly the same as introduced in the original ResNet paper [40].

V. Results

In this section, prediction methods are tested using the six basic and neutral expressions.
The RF and the ResNet algorithm used here are not deterministic. Therefore, all results are presented with upper and lower bound of a 95% confidence interval from 30 tests.

In the Expressions section of Tables I, II, IV, and V. *An* represents anger, *Di* represents disgust, *Fe* represents fear, *Ha* represents happy, *Ne* represents neutral, *Sa* represents sadness and *Su* represents surprise expression. Also at the bottom of these tables, * refers to the mean accuracy across all categories.

TABLE I
PERCENTAGE OF CORRECTLY CLASSIFIED CLASSES BY RESNET AND THE PROPOSED METHOD WITHOUT ADVERSARIAL ATTACK

Attacks	Method	**An**	**Di**	**Fe**	**Ha**	**Ne**	**Sa**	**Su**
None	ResNet	98	96	100	100	100	88	100
	Proposed	98	97	97	99	96	97	96

*$ResNet$ =97.43%±0.1, Proposed = 97.14%±0.1

Table I shows the accuracy obtained by ResNet and the proposed method on each class of the CK+ database with no adversarial attack applied on the database. The proposed method and the ResNet model all achieved not less than 88% accuracy across all classes. Also, while the ResNet model outperformed the proposed method with a small margin of 0.29% in the overall accuracy achieved in this case, $t(59) = 0.24$, $p = .21$ of two-sample t-test shows that the difference was not significant (*ResNet M = 97.43, proposed M = 97.14*).

Table II shows the accuracy obtained by ResNet and the proposed method after *Type A* adversarial attack is applied to the anger class of the training data.

TABLE II
PERCENTAGE OF CORRECTLY CLASSIFIED CLASSES BY RESNET AND THE PROPOSED METHOD AFTER TYPE A ADVERSARIAL ATTACK IS LAUNCHED ON ANGER CLASS

Attacks	Method	**An**	**Di**	**Fe**	**Ha**	**Ne**	**Sa**	**Su**
Type A	ResNet	4	98	80	96	96	94	98
	Proposed	96	100	92	100	98	94	98

*$ResNet$ =80.86%±0.3, Proposed = 96.86%±0.2

Although the accuracy achieved by the proposed method experienced a decrease (from 98% to 96%) for the anger

class after the attack is launched, the ResNet model sees almost all the images in the anger class as different after the attack and therefore failed to recognize most of the anger images correctly, which results in only 4% recognition accuracy achieved for the anger class.
This has a strong impact on the overall accuracy achieved by the ResNet model. As can be seen, even though the accuracy achieved by both ResNet and the proposed method decreases after the attack, the proposed method only experienced a small decrease of 0.28% compared to the ResNet model that experiences a large decrease of up to 16.57%.

TABLE III
ACCURACY OBTAINED FROM PREDICTIONS OF NORMAL IMAGES WHEN CERTAIN CLASSES ARE TRAINED WITH TYPE A ADVERSARIAL IMAGES

Class trained with type A	**ResNet**	**Proposed**
Anger	80.86±0.3	96.86±0.2
Disgust	75.43±0.3	97.28±0.1
Fear	81.14±3.0	97.0±0.1
Happy	77.43±0.3	97.28±0.2
Neutral	74.86±0.3	96.86±0.1
Sad	78.57±1.3	96.86±0.2
Surprise	81.43±0.3	96.86±0.1

Table III shows the accuracy obtained by ResNet and the proposed method when *Type A* adversarial attack is launched on different classes of the CK+ database. Here, the ResNet model has shown a high recognition accuracy when the attack is launched on fear and surprise classes with an achieved accuracy of up to 81.14% and 81.43% respectively. On the other hand, the proposed method has shown a high recognition accuracy when the attack is launched on disgust and happy classes with an accuracy of 97.28% in both cases.

TABLE IV
PERCENTAGE OF CORRECTLY CLASSIFIED CLASSES BY RESNET AND THE PROPOSED METHOD AFTER TYPE B ATTACK IS LAUNCHED ON THE DATA

Attacks	Method	**An**	**Di**	**Fe**	**Ha**	**Ne**	**Sa**	**Su**
Type B	ResNet	90	92	84	94	92	84	100
	Proposed	96	100	92	100	100	90	100

*$ResNet$ =90.86%±1.5, Proposed = 96.86%±0.2

Table IV shows the percentage accuracy obtained by ResNet and the proposed method after the *Type B* adversarial attack is launched on all test images. Surprisingly, although the accuracy of the ResNet model is reduced in each class after the attack has been launched, the accuracy remains the same on the surprise class even after the attack. The reason behind that might be that the extracted features of images in the surprise class before the attack more or less resemble the extracted features even after the attack has been launched. Nevertheless, the attack still reduces the overall accuracy of the ResNet model to 90.86%.

Conversely, the proposed method manages to achieve 100% on four different classes (disgust, happy, neutral, and surprise) even though the method is not trained with any of the adversarial images.

The proposed method achieves an overall accuracy of up to 6% higher than the ResNet model.

TABLE V
PERCENTAGE OF CORRECTLY CLASSIFIED CLASSES BY RESNET AND THE PROPOSED METHOD AFTER FSGM ADVERSARIAL ATTACK IS LAUNCHED ON THE DATA

Attack	Method	**An**	**Di**	**Fe**	**Ha**	**Ne**	**Sa**	**Su**
FGSM	ResNet	84	90	92	90	90	86	98
	Proposed	98	100	94	100	100	96	94

*$ResNet$ =90.0%±0.66, Proposed = 97.43%±0.2

Table V shows the percentage accuracy obtained by ResNet and the proposed method on each class after the FGSM adversarial attack is applied to the test data. The accuracy achieved in each class reduces compared to the result in Table I when no adversarial attack is applied to the images. The anger class experienced the most decrease of up to 14% followed by the happy and neutral classes that experienced a decrease of 10%.

However, an accuracy of up to 100% is achieved in three classes (disgust, happy, and neutral) when the same data in which the FGSM method is applied is predicted by the proposed method. Also, the accuracy achieved by any one class by the proposed method surpasses the accuracy achieved in all the classes when the prediction is made by the ResNet model except for the surprise expressions.

As the ϵ of 0.01 is so small, the resulting adversarial image remains the same for the human eye even after the FGSM method is applied to the image. However, the generated adversarial image still manages to fool the ResNet model and reduces the overall accuracy achieved by the model with more than 7%.

Across all techniques, the time it takes to evaluate the result by both methods is considerably lower in our case when compared to the ResNet model. While our method takes between 14.8±0.2 secs to to 1.4±0.3 mins to execute in different cases, it takes between 32.83± 0.3 mins to 5.48± 0.5 hrs for the ResNet algorithms to execute on the same machine.

Also, a regression analysis was used to check the effectiveness of the proposed method with respect to the ResNet model in all scenarios after the attack, $p < .001$ of Ordinary Least Squares (OLS) method shows that the difference is significant.

VI. DISCUSSION

Initially, while the ResNet model outperformed the proposed method by 0.29% before the adversarial attack has been launched on the images, the proposed method outperformed the ResNet model in all cases when *Type A*, *Type B*, and *FGSM* adversarial attacks are applied to the database. This is due to the fact that the deep model uses the color distribution rather than the image pattern when fed directly

on raw images. As a result, the deep model is not able to efficiently analyze emotion labels when certain changes are applied to the kind of images the deep model is trained on.

Although different people might show different facial displays when experiencing the same kind of emotion, a model trained to recognize emotion should be able to perform with a little or no reduction in accuracy when tested with the same images with a small change, unlike the reduced accuracy of up to 22% of the ResNet model in Section V.

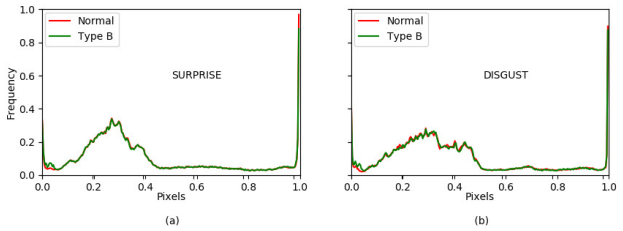

Fig. 7. Histogram showing the color distribution of (a) surprise and (b) disgust class before and after Type B adversarial attack is launched on the data

As can be seen from Table IV and Table V, we can say the ResNet model is good in predicting surprise expression among other expressions when *Type B* and the FGSM adversarial attack is applied on the data, achieving an accuracy of 100% which is equal to the achieved accuracy by the proposed method on the same data and an accuracy of 4% more than the proposed method when FGSM adversarial attack is launched on the data. Since the ResNet model extracts the features based on the color distribution of the images, the reason behind that could be the color distribution for the surprise class before and after the attack is launched on the data is more or less the same. Fig. 7 is a histogram showing the color distribution on images in surprise and disgust expressions before and after *Type B* adversarial attack is launched on the data. As can be seen, the distribution is almost the same across the whole histogram in surprise expressions compared to when the same attack is launched on disgust expression.

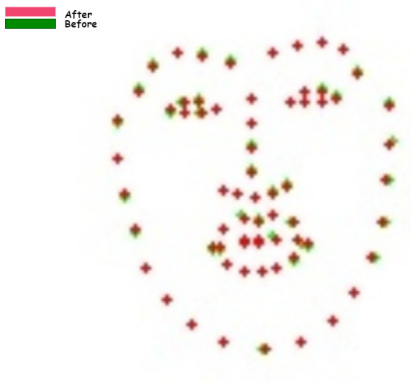

Fig. 8. Image showing extracted landmarks by the proposed method before and after the FGSM adversarial attack is launched on the data

Similarly, we can say that the features extracted by the proposed method across the whole dataset before and after the attack are more or less the same as we have observed the same or small decrease in the results achieved by the proposed method in Section V. Fig. 8 shows sample landmarks extracted by the proposed method before and after the FGSM adversarial attack is launched on a particular image. Almost all the landmarks extracted after the attack are placed on the same point with the landmarks extracted before the attack when observed in the figure.

Considering the difference between the extracted features of the test data before and after the FGSM adversarial attack is launched on the data, we would expect that the difference between these two features should be somewhere close to zero as evidence shows that the landmarks extracted by the proposed method before the attack are more or less the same even after the attack since the method is not very sensitive to the attack. As expected, the difference between the features is located around zero-axis in both x and y coordinates as shown in Fig. 9.

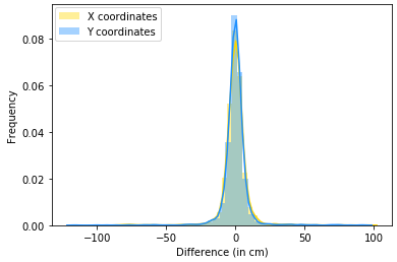

Fig. 9. Histogram showing the difference between the features extracted by the proposed method before and after the FGSM adversarial attack is launched on the data

In a few cases, the result obtained by the proposed method after the application of the adversarial attack is even higher than the result obtained before the adversarial attack is launched on the database. This is due to the nondeterministic nature of the RF algorithm. But by looking at the upper and lower bound of the accuracies in Table III, the results are comparable.

In addition, we checked the significance of the result obtained by the proposed method and the ResNet model before the application of the adversarial attack, but no significant difference was found. However, the difference was found to be significant in all the different scenarios after the attack.

VII. CONCLUSIONS

This paper proposed an adversarial attack resistant based approach to analyze emotion in images of faces using landmarks. Our method outperformed the ResNet model in classifying images in various cases after an adversarial attack has been launched on the data. Furthermore, the proposed method achieved a comparable result, with no significant difference to the ResNet model even when no adversarial attack was applied to the data.

This study is carried out on a single database. Future work should apply the same approach to different databases to test generalizability.

REFERENCES

[1] Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(6), 1113–1133. https://doi.org/10.1109/TPAMI.2014.2366127

[2] Olivares-Mercado, J., Toscano-Medina, K., Sanchez-Perez, G., Portillo-Portillo, J., Perez-Meana, H., & Benitez-Garcia, G. (2019). Analysis of hand-crafted and learned feature extraction methods for real-Time facial expression recognition. 2019 7th International Workshop on Biometrics and Forensics, IWBF 2019, 1–6. https://doi.org/10.1109/IWBF.2019.8739178

[3] Mehrabian, A. (1981) "Silent Messages"-A Wealth of Information About Nonverbal Communication (Body Language).

[4] Jameel, R., Singhal, A., & Bansal, A. (2016). A comprehensive study on Facial Expressions Recognition Techniques. Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering, Confluence 2016, 478–483. https://doi.org/10.1109/CONFLUENCE.2016.7508167

[5] Schmelzer, R., (2020). Can AI Detect Your Emotion Just By How You Walk? in Forbes Web Address: https://www.forbes.com/sites/cognitiveworld/2020/03/29/can-ai-detect-your-emotion-just-by-how-you-walk/#74ebee3b69de

[6] Westermeyer, J. (1979). A social interactional theory of emotions. Amer. J. Psychiatry, vol. 136, no. 6, pp. 870–870.

[7] Freiwald, W. A., Tsao, D. Y. & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe, Nature Neuroscience., vol. 12, no. 9, pp. 1187–1196.

[8] Gui, J., Zhang, Y., Li, S., Xu, P., & Lan, S. (2016). Real-Time 3D Facial Subtle Expression Control Based on Blended Normal Maps. Proceedings - 2015 8th International Symposium on Computational Intelligence and Design, ISCID 2015, 1, 466–469. https://doi.org/10.1109/ISCID.2015.200

[9] Gross, M., Gertsner, G., Koditschek, D. E., Fredrickson, B. L., & Crane, E. A. (2006). Emotion Recognition from Body Movement Kinematics. Journal of Chemical Information and Modeling, 53, 160. https://doi.org/10.1109/ACCESS.2019.2963113

[10] Longmore, C. A., & Tree, J. J. (2013). Motion as a cue to face recognition: Evidence from congenital prosopagnosia. Neuropsychologia, 51, 864-875

[11] Hancock, P. (2008). Psychological image collection at Stirling (PICS)," Web address: http://pics.stir.ac.uk/ESRC/index.htm

[12] Mollahosseini, A., Hasani, B., Mahoor, M. H., (2017). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. in IEEE Transactions on Affective Computing.

[13] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee D.-H. (2015). Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63.

[14] Izquierdo-Reyes, J., Ramirez-Mendoza, R.A., Bustamante-Bello, M.R., et al. (2018). Emotion recognition for semi-autonomous vehicles framework. Int J Interact Des Manuf 12, 1447–1454. https://doi.org/10.1007/s12008-018-0473-9

[15] Wu, J., & Fu, R. (2019). Universal, transferable and targeted adversarial attacks. Retrieved from http://arxiv.org/abs/1908.11332

[16] Harding, S. M., Rajivan, P., & Bertenthal, B. I. (2016). Human Decisions on Targeted and Non-Targeted Adversarial Samples. (August), 451–456.

[17] Ekman, R. (1997). What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA.

[18] Boyko, N., Basystiuk, O., & Shakhovska, N. (2018). Performance evaluation and comparison of software for face recognition, based on dlib and OpenCV library. In 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP) (pp. 478-482). IEEE.

[19] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, (May 2014), 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

[20] Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. Proceedings - 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000, (March), 46–53. https://doi.org/10.1109/AFGR.2000.840611

[21] Ekman, P. and Friesen, W. V., (1971). Constants across cultures in the face and emotion. Journal of personality and social psychology, vol. 17, no. 2, p. 124.

[22] Akyol, F. and Şahin, P. D., (2016). Image-based facial expression detection, 24th Signal Processing and Communication Application Conference (SIU), Zonguldak, 2016, pp. 609-612. https://doi.org/10.1109/SIU.2016.7495814

[23] Wang, Y., Li Y., Song, Y. & Rong, X. (2019). The Application of a Hybrid Transfer Algorithm Based on a Convolutional Neural Network Model and an Improved Convolution Restricted Boltzmann Machine Model in Facial Expression Recognition, in IEEE Access, vol. 7, pp. 184599-184610. https://doi.org/10.1109/ACCESS.2019.2961161

[24] Melaugh, R., Siddique, N., Coleman, S., & Yogarajah, P. (2019). Facial expression recognition on partial facial sections. International Symposium on Image and Signal Processing and Analysis, ISPA, 2019-September, 193–197. https://doi.org/10.1109/ISPA.2019.8868630

[25] Lyons, M. J., Budynek, J. & Akamatsu, S. (1999) Automatic classification of single facial images. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 12, pp. 1357–1362.

[26] Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

[27] Xu, M., Cheng, W., Zhao, Q., Ma, L., & Xu, F. (2016). Facial expression recognition based on transfer learning from deep convolutional networks. Proceedings - International Conference on Natural Computation, 2016-January, 702–708. https://doi.org/10.1109/ICNC.2015.7378076

[28] Zhang, X., Zhang, L., Wang, X. J., & Shum, H. Y. (2012). Finding celebrities in billions of web images. IEEE Transactions on Multimedia, 14(4 PART1), 995–1007. https://doi.org/10.1109/TMM.2012.2186121

[29] Tian, Y., Cheng, J., Li, Y., & Wang, S. (2019). Secondary Information Aware Facial Expression Recognition. IEEE Signal Processing Letters, 26(12), 1–1. https://doi.org/10.1109/lsp.2019.2942138

[30] Li, S., Deng, W., & Du, J. (2019). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing, 28(1), 356-370. https://doi.org/10.1109/TIP.2018.2868382

[31] Lu, Y., Wang, S., Zhao, W., & Zhao, Y. (2019). WGAN-Based Robust Occluded Facial Expression Recognition. IEEE Access, 7, 93594–93610 https://doi.org/10.1109/ACCESS.2019.2928125

[32] Dlib Python API Tutorials [Electronic resource] – Access mode: http://dlib.net/python/index.html

[33] Goodfellow I. J., Shlens J., & Szegedy C. (2014). Explaining and Harnessing Adversarial Examples. CoRR abs/1412.6572 (2014). http://arxiv.org/abs/1412.6572

[34] Papernot, N., Faghri, F., Carlini, N., Goodfellow I., Feinman R., Kurakin A., Xie C., Sharma Y., Brown T., Roy A., Matyasko A., Behzadan V., Hambardzumyan K., Zhang Z., Juang Y., Li Z., Sheatsley R., Garg A., Uesato J., Gierke W., Dong Y., Berthelot D., Hendricks P., Rauber J. & Long R. (2018). Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. arXiv preprint arXiv:1610.00768

[35] Tran, E., Mayhew, M. B., Kim, H., Karande, P. & Kaplan A. D., Facial Expression Recognition Using a Large Out-of-Context Dataset. 2018 IEEE Winter Applications of Computer Vision Workshops (WACVW), Lake Tahoe, NV, 2018, pp. 52-59. http://doi: 10.1109/WACVW.2018.00012

[36] Simonyan, K. & Zisserman, A. (2015) Very deep convolutional networks for large-scale image recognition. In ICLR, arXiv preprint arXiv:1409.1556

[37] Shehu, H. A., Tokat, S., Sharif, M. H., Uyaver, S. (2019). Sentiment analysis of Turkish Twitter data. American Institute of Physics (AIP) Conference Proceedings, 080004(December). https://doi.org/10.1063/1.5136197

[38] Breiman, L. (2001). Random Forests. Machine Learning. 45 (1): 5–32. doi:10.1023/A:1010933404324

[39] Kaminski, B.; Jakubczyk, M.; Szufel, P. (2017). A framework for sensitivity analysis of decision trees. Central European Journal of Operations Research. 26 (1): 135–159. doi:10.1007/s10100-017-0479-6

[40] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 770–778. https://doi.org/10.1109/CVPR.2016.90