An Out-of-distribution Attack Resistance Approach to Emotion Categorization

Harisu Abdullahi Shehu, *Graduate Student Member, IEEE*, Will N. Browne, *Member, IEEE*, and Hedwig Eisenbarth

***Abstract*—Deep neural networks are a powerful model for feature extraction. They produce features that enable state-of-the-art performance on many tasks, including emotion categorization. However, their homogeneous representation of knowledge has made them prone to attacks, i.e. small modification in train or test data to mislead the models. Emotion categorization can usually be performed to be either in-distribution (train and test with the same dataset) or out-of-distribution (train on one or more dataset(s) and test on a different dataset). Our already developed landmark-based technique, which is robust for in-distribution improvement against attacks in emotion categorization, could translate to out-of-distribution classification problems. This is important as different databases might have different variations such as in color or level of expressiveness of emotion. We compared the landmark-based method with four state-of-the-art deep models (EfficientNetB0, InceptionV3, ResNet50 and VGG19), as well as emotion categorization tools (i.e. Py-Feat and the Azure Face API) by performing a cross-database experiment across six commonly used databases, i.e. CK+, JAFFE, KDEF, NIMH-ChEF, RAF, and PICS databases. The landmark-based method has achieved a significantly higher accuracy, achieving an average of 47.44% compared with most of the deep networks ($< 36\%$) and the emotion categorization tools ($<37\%$) with considerably less execution time. This highlights that out-of-distribution emotion categorization is a much harder task due to detecting underlying emotional cues than emotion categorization in-distribution where superficial patterns are detected to $> 97\%$ accuracy.**

***Impact Statement*—Recognising emotions from people's faces has real-world applications for computer-based perception as it is often vital for interpersonal communication. Emotion recognition tasks nowadays are addressed using deep learning models that model colour distribution so classify images rather than emotion. This homogeneous knowledge representation is in contrast to emotion categorization, which is hypothesised as more heterogeneous landmark-based. This is investigated through out-of-distribution emotion categorization problems, where the test samples are drawn from a different dataset to training images. Our landmark-based method achieves a significantly higher classification performance (on average) compared with four state-of-the-art deep networks (EfficientNetB0, InceptionV3, ResNet50 and VGG19), as well as other emotion categorization tools such as Py-Feat and the Azure Face API. We conclude that this improved generalization is relevant for future developments of emotion categorization tools.**

***Index Terms*—Attack, Emotion categorization, Emotion recognition, Facial expression, Facial landmarks, Cross-database, Out-of-distribution**

I. Introduction

EMOTION categorization is an important task in understanding how human beings convey their emotional state. It is one of the most widely studied fields in human-computer interaction [1].

Nowadays, studies have been conducted on emotion categorization because of its importance for sociable robots [2]. For instance, the increasing demand to introduce robots to help people with disabilities walk again [3] has made it important to recognize an emotional state of mind as these robots need to understand people's emotions to be able to interact in an intuitive way.

Emotion categorization is generally performed in two different ways; i) in-distribution, which performs training and testing on the same dataset [4], and ii) out-of-distribution, in which the training and the test data comprises of one or more dataset, with no overlap [5].

Several methods have been developed to categorize emotion from images, most of which perform in-distribution classification (within the same dataset). However, performing in-distribution classification does not ensure generalization in emotion categorization as the trained classifier is tested with similar images from the same image distribution. Moreover, real-world data is not within the same distribution when we meet new people or new environments.

Deep learning (DL) algorithms have powerful feature learning abilities and have been used by various researchers to perform end-to-end learning in recent years [6], [7]. However, despite being a powerful feature learning tool, DL algorithms are strongly affected by high inter-subject variations that exist due to attributes such as race, level of expressiveness, and ethnicity, etc., [8], [9] that are non-linearly co-founded with emotion categorization.

To address these issues, we have developed a landmark-based technique that has shown robustness to changes such as color and noise distortion for in-distribution emotion categorization [10]. However, it is unknown if the technique translates to out-of-distribution (cross-database classification) as different databases might vary in terms of the illumination, age, the ethnicity of subjects, as well as the level of expressiveness of the emotion being expressed.

This paper is an expanded paper published in the proceedings of the IEEE RO-MAN Conference, which was held (in Naples, Italy) virtually as live interactive conference between 31 August - 4 September, 2020.
This work was not funded by any organization.
Manuscript received February 8, 2021; revised June 15, 2021.

Harisu Abdullahi Shehu is with School of Engineering and Computer Science, Victoria University of Wellington, 6012 Wellington, New Zealand (e-mail: harisushehu@ecs.vuw.ac.nz).

Will N. Browne is with School of Electrical Engineering and Robotics, Queensland University of Technology, 4000 Brisbane, Australia (e-mail: will.browne@qut.edu.au).

Hedwig Eisenbarth is with School of Psychology, Victoria University of Wellington, 6012 Wellington, New Zealand (e-mail: hedwig.eisenbarth@vuw.ac.nz).

This paragraph will include the Associate Editor who handled your paper.

This research aims to analyze whether a landmark-based method will be robust in out-of-distribution emotion categorization. A novel method will be used to help with head recognition and alignment. The performance will be compared with four state-of-the-art DL algorithms; EfficientNetB0 [11], InceptionV3 [12], ResNet50 [13], VGG19 [14] models, as well as emotion categorization tools such as Python Facial Expression Analysis Toolbox (Py-Feat) [15], and the commercial emotion categorization model, i.e. the Microsoft Azure Face Application Programming Interface (Azure Face API) [16] to analyze which of the techniques will achieve a better performance.

The cross-database evaluation will use six state-of-the-art (CK+ [17], KDEF [18], JAFFE [19], NIMH-ChEF [20], RAF [21], and PICS [22]) emotion databases. Most of the publicly available datasets consist of acted facial expressions, which are biased by authors' design decisions. In order to take this variability into account, we use several databases in the training process. With a leave-one dataset-out approach, we will train interactively on five different datasets and test on the remaining dataset.

Support Vector Machines (SVM) [23] will be used as the classifier to categorize the extracted landmarks in order to automatically capture the complex relationship between data points. The final objective is to analyze the difference in complexity between the in and out-of-distribution tactics.

The rest of the paper is organized as follows: Section II highlighted recent work carried out on emotion categorization using deep learning. The section also provides the required background knowledge of facial landmarks and machine learning. Section III explains how facial landmarks are processed after extracting them from images. The section also explains the possible problems of using facial landmarks to categorize emotion and how these problems are addressed. Section IV introduces the datasets used, how the experiment is set up, as well as presents the experimental results. Section VI provides a further discussion on the obtained result and Section V concludes the paper and hints at further studies.

II. BACKGROUND

The goal of this section is to highlight the recent in uential work performed on emotion categorization and provide knowledge on facial landmarks and the machine learning techniques used in this research.

A. Deep Learning-based Approach to Emotion Categorization

There have been several methods developed to categorize emotion from images. One of the most commonly used methods to categorize emotion from images is the use of convolutional neural networks.

Videla et al. [24] proposed a 10-layer convolutional neural network (CNN) to classify facial expression from images of CK+ and the JAFFE database. Faces in the images were initially detected, cropped, resized, and then fed to the proposed CNN to perform end-to-end learning. The proposed CNN achieved an accuracy of 99.3% on the CK+ and 78.1% on the JAFFE database. The study only used the last-three frame images from the CK+ database. However, as the CK+ is a database of posed[1] facial expression starting from neutral to peak frames, i.e. where the emotion is expressed at the highest intensity, it is unknown if the method will achieve a high classification accuracy on the mid-frames where the emotion is not expressed at peak. Besides, peak emotions are rare in real-life.

Sokolov et al. proposed a CNN architecture [25] to categorize emotion from a cross-platform application in real-time. The application was developed to categorize facial expressions captured with a frontal camera. Emotions were estimated in the arousal-valence scale, i.e. how valence or aroused person is. The training was done on datasets that were manually assembled from open-source data. The developed CNN achieved an accuracy of 64.89% on the validation set and 63.01% on the test set. Considering that the application was developed to categorize emotion based on two different classes categorizing the emotion intensity to be either high/low valence or high/low arousal, the achieved 63% accuracy is a little bit better than random guessing, which in this case is 50%. As such, questions remain as to whether the developed application will perform well on the commonly used six basic (or the six basic plus neutral) emotional expressions [26].

Verma et al. proposed a convolution neural architecture called the Venturi architecture [27] to categorize the six basic plus neutral expressions from images of the KDEF database. The performance of the model is compared with rectangular architecture and the modified triangular architecture [28]. The Venturi architecture achieved an accuracy of 86.78% on the KDEF database compared with the 79.61% and 82.7% accuracy achieved by the rectangular and modified triangular architecture. The proposed Venturi architecture achieved an accuracy of 98.87% on the training set. Hence, there is a huge difference between the accuracy of the training and test sets (> 12%). Therefore, questions remain as it is unclear if the model overfits the data on the test set.

DL algorithms have been used [24], [25], [27] to categorize emotion from images. However, feeding DL algorithms with face images to perform end-to-end learning analyzes the color distribution of the pixels in an image, which may not generalize in a cross-database emotion categorization task as different databases might have varying color distribution of their pixels. Besides, expression of emotion might not be based on color alone.

B. In-distribution Approach to Emotion Categorization

Tang et al. [4] proposed two versions of a frequency-based neural network approach called the Basic frequency neural network (Basic-FreNet) and Block frequency neural network (Block-FreNet) to categorize emotion in the CK+ and KDEF datasets. In the Basic-FreNet, a learnable multiplication kernel was applied to learn features in the frequency domain followed by a summarizing layer which yields high-level features. In Block-FreNet, the weight-shared kernel was designed for

[1] The posed dataset is the type of dataset captured in a controlled environment based on instructions given by an experimenter.

feature learning and dimension reduction. Experimental results showed that the Block-FreNet achieved up to 98.91% and 91.22% accuracy on the CK+ and the KDEF dataset. However, the experiment was conducted separately on the CK+ and KDEF databases. Questions remain if the method will achieve higher accuracy on either of the datasets if the training was conducted on a different dataset.

Recently, we developed a method to categorize emotion from images using facial landmarks [10] as this was hypothesized to better represent emotion encoding than colour distribution. Initially, the facial landmarks were detected using the dlib [29] library. Detected landmarks were pre-processed and used to categorize images from the CK+ database. The method was tested with adversarial images and compared to a state-of-the-art deep learning algorithm, i.e. the ResNet model. The developed method has been demonstrated to be more robust against changes to images than the ResNet model. We conducted the research to investigate robustness against changes made to (in-distribution) images of a single database. However, it is unknown if the method will be robust to categorize images from different datasets as different datasets might have a different variation in terms of the way the emotion is expressed.

Several methods have been developed to categorize emotion from images of a single dataset [4], [10] . However, performing a self-classification within the same dataset in emotion categorization does not ensure generalization since the same dataset generally consists of similar images with the same level of expressiveness by the participants, unlike in other databases, in which a very high variation in the level of expressiveness of the emotion might exist.

C. Out-of-distribution Approach to Emotion Categorization

Mayer et al. [5] developed a facial expression recognition method by fitting a 3D model onto faces of emotional images. Thereafter, the developed method was used for cross-database classification [30] to classify images from the CK+, MMI [31], and FEEDTUM [32] databases. Regardless of the number of images from the databases, Mayer et al. randomly extracted two subsets of images with equal sizes from each database. One of the sets was used for training whereas the other set was used to test the performance of the classifier. The approach achieved higher or comparable results across all the datasets. Considering that only a certain part of the images was extracted from each dataset to perform the classification, questions remain as to whether the method will still perform well on the complete dataset.

Li et al. [33] proposed a deep learning-based approach called the Deep Emo-transfer Network (DETN) to analyze cross-database recognition of facial expressions from images. The proposed DETN extends the work of Long et al. [34] by including a multi-kernel maximum mean discrepancy (MK-MMD). Li et al. believed that the possible bottleneck for cross-domain expression is due to the skew in classification distribution between the source and the target domain and therefore introduce a learnable class-wise weighting parameter to the original MMD. During the experiment, the RAF database was used for training, and evaluation was made on the CK+, JAFFE, MMI, SFEW [35], and FER2013 [36] dataset. The method was found to achieve a better result than many state-of-the-art methods for this cross-database experiment. Li et al. trained their method only on the RAF database. It is unclear whether the method will achieve higher performance if the training was performed on a different dataset or a group of datasets. Besides, it is beneficial to train on a variety of different data as the real-world data variations are unlikely to be captured in a single dataset.

Certain methods have been used to perform cross-database experiments [5], [33]. However, these methods use only a selected portion of the data or train on a single dataset. Considering that most databases are biased by authors' design decisions, it is unclear if these methods will generalize well when training is performed on other state-of-the-art databases or complete data rather than using only a subset of images from each dataset.

This research is needed to analyze whether a landmark-based method will be insensitive to changes such as the level of expressiveness across different databases and thereby translate to an out-of-distribution emotion categorization.

D. Facial Landmarks

The facial landmarks are the location of (x, y) coordinates that maps the facial structures on the face. They locate and represent salient regions of the face such as the eyes, nose, mouth, eyebrows, and jawline. There are several methods such as i) 194-point and ii) 68-point models, etc. that can be used to detect facial landmarks. The 194-point model can be trained on the HELEN dataset [37] whereas the 68-point model can be trained on the iBUG 300-W dataset [38]. However, as higher landmark detection accuracy was shown on the 300W benchmark [39] using the 68-point landmark, in this research, the pre-trained landmark detector in the dlib [29] library is used to detect the 68-point landmark coordinates.

E. Machine Learning

1) SVM: Support Vector Machines (SVM) are a supervised machine learning algorithm that finds hyperplane(s) in an N-dimensional space (where N is the number of features). The hyperplanes are decision boundaries that help to classify data points.

SVM is chosen to be used in this research as the algorithm has the capability to automatically capture a complex relationship between data points without having to perform a difficult data transformation manually.

2) Deep Learning: Deep learning (DL) algorithms are a powerful method for feature learning and extraction. The algorithms transform low-level input data into a high-level abstraction in complex data [40]. EfficientNetB0 [11], InceptionV3 [12], ResNet50 [13], and VGG19 [14], were chosen to be used in this research as they are one of the most commonly used deep networks used for classification problems [41].

ResNet50 was chosen to be used in this research as it uses skip connections and adds zero new parameters to the existing ones, which makes it faster to train compared to other DL algorithms.

VGG19 was chosen to be used in this research as it is the latest version of the VGG models and has more weight layers (19) compared to other VGG models such as the VGG16 with 16 weight layers, which should add greater delity to the learnt features.

InceptionV3 was chosen as it provides several improvements such as the use of auxiliary classi er and factorization of 7 7 convolutions, etc. to its previous versions (i.e. Inception version 1 and 2). As a result, the network is faster and lighter compared to the previous versions.

Finally, Ef cientNetB0 was chosen to be used as it uses a simple yet effective compound coef cient to scale up a CNNs dimension in a more structured manner. Thereby improving accuracy and ef ciency.

F. Emotion Categorization Tools

Two emotion categorization tools, i.e. Azure Face API and Py-Feat were chosen to be used in this research.

Azure Face API was chosen to be used as it is from a well-established vendor (i.e. Microsoft) who have a large amount of resources to train deep networks and other algorithms on this problem.

Py-Feat [15] was chosen to be used as it is a recently introduced toolbox, which has achieved a higher performance result compared to other emotion categorization frameworks such as the DeapFace [42] on commonly used state-of-the-art (e.g. AffectNet [43]) emotion categorization dataset.

1) Azure Face API: The Azure Face service is a commercial model developed by Microsoft that uses machine learning to perform operations, including emotion categorization on human faces in images[16].

Currently, there are three recognition models available in the Azure Face service; the recognition_01 published in 2017, the recognition_02 published in 2019, and the recognition_03 model which was published in 2020. These models are continually being improved based on customer feedback and advances in research. However, the recognition_03 model is currently the most accurate model available [44] and is therefore recommended to be used by Microsoft. For that reason, in this research, the recognition_03 model will be used to categorize emotion of the six datasets used in this research.

2) Python Facial Expression Analysis Toolbox (Py-Feat): The Py-Feat is an open-source facial analysis toolbox, including emotion categorization. The model is developed to help domain experts disseminate and benchmark their facial expression analysis models. For that reason, we chose to use the model to benchmark our method.

At the moment, there are four different Py-Feat models, which includes residual masking network model (resmasknet), a multi-layer conventional neural networks (feat-ferNet), random forests (feat-rf), and linear SVM (feat-svm) that can be used for emotion categorization from images. However, in this research, the default model, i.e. resmasknet was chosen to be used as it yielded the best performance results compared with the other models.

III. METHOD

A. Pre-processing

Emotion labels were converted to integers and then one-hot encoded. We also converted the image pixels to an array and then normalized the pixel values to a range [0, 1] to enable fast computation.

B. Feature Extraction

To enable a better out-of-distribution performance, certain improvements have been made to the proposed method in our previous research [10]. Algorithm 1 presents the procedure of extracting facial features from emotional images. In contrast to our previous research that detected faces in an image with the help of the Dlib frontal face detection model, here, faces are detected in two different ways. Initially, faces in an image are detected with the help of the CAFFE model [45] as it is more reliable than other face detection models such as the Haar cascade (HC) model [46]. However, if the face detection con dence obtained by the CAFFE model is $< 50\%$, it is assumed that the face may not be accurately located and therefore the HC model is utilized to locate the faces in the images.

Thereafter, we detect the $x, (y)$ landmark coordinates from faces using the dlib library as shown in Fig. 1.

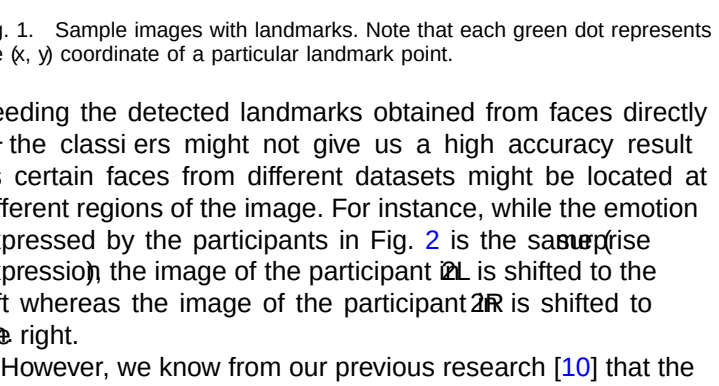

Fig. 1. Sample images with landmarks. Note that each green dot represents the (x, y) coordinate of a particular landmark point.

Feeding the detected landmarks obtained from faces directly to the classi ers might not give us a high accuracy result as certain faces from different datasets might be located at different regions of the image. For instance, while the emotion expressed by the participants in Fig. 2 is the same surprise expression, the image of the participant in 2L is shifted to the left whereas the image of the participant in 2R is shifted to the right.

However, we know from our previous research [10] that the relationship between each landmark coordinate is usually the same or relatively similar to one another in a given emotion. Therefore, we de ne the $x, (y)$ landmark coordinates of the tip of the nose as the center of the face. Consequently, we nd the distance of each landmark coordinate from the center of the face.

Moreover, as certain faces might be tilted (see Fig. 3), we correct that by assuming that the nasal bridge is straight for all participants in the databases. Therefore, we calculated the angles of the nose bridge and offset all calculated angles by the same angle.

Algorithm 1: Procedure adopted by the landmark-based method

```
Description: Extracting facial features from emotional
             images for emotion categorization
Input  : Test images L_ti from all seven
         categories, i.e anger, disgust, fear,
         happy, neutral, sad, surprise
Output : x, y, xDistance, yDistance, offset
Initialize
L_ti  List of images
for Image i in L_ti do
    flag = false
    confidence = CAFFE Detect Face(i) % detect face
     using caffe model
    if confidence < 0.5 then
        %confidence not reliable
        flag = true
    end
    if flag == true then
        HC Detect Face(i) % detect face using Haar
         cascade classifier
    end
    detect landmark (face) %detect (x, y) coordinates
     of facial landmarks
    for i in range(0, 68) do
        if i == 33 then
            % landmark of tip of the nose
            xCenter, yCenter = landmark(i) %assign
             landmark of tip of the nose as the center
             of face
        end
    end
    for each (x,y) in detected_landmarks do
        xDistance = x - xCenter
        yDistance = y - yCenter
        offset = ((y;x) 360 / (2 ))
        append x, y, xDistance, yDistance, and offset to
         vector
    end
end
```

Fig. 2. Sample emotion expressed at different locations of images from the CK+ database.

Fig. 3. Adapted expressions with tilted faces. Note that the first, second, and third images (from the left) represent sample expressions from JAFFE, NIMH-ChEF, and the CK+ database respectively.

Finally, we fed the classifier with detected (x,y) landmark coordinates, the distance of each point from the center, as well as the calculated angle to classify the emotion images into their respective categories.

IV. EXPERIMENTAL WORK

A. Hardware specification

A 24GB Graphical Processing Unit (GPU) device Quadro RTX 6000 with CUDA version 10.2 is used in this research.

B. Data sets

1) CK+: The Extended Cohn-Kanade (CK+) dataset [17] is a dataset of posed facial expressions of 201 adults between the ages of 18-50 years. 13% of the participants in the database are Afro-American, 81% are Euro-American, and the remaining 6% are from other groups.

This research uses the last-half frames of the CK+ database based on the technique developed by Shehu et al.[47], [48] as it has been shown to give a higher accuracy result. Overall, we used a total of 3,368 images that consists of the six basic plus neutral expression. The first row of Fig. 4 shows examples of expressions from the CK+ database.

2) JAFFE: The Japanese Female Facial Expression (JAFFE) [19] is a posed facial expression database of 10 Japanese models. The database has a total of 213 grey images that consists of the six basic plus neutral expressions.

All 213 images from the JAFFE database were used in this research. The fifth row of Fig. 4 shows examples of expressions from the JAFFE database

3) KDEF: The Karolinska Directed Emotional Faces (KDEF) [18] is a dataset of posed facial expressions of 70 (35 male and female) amateur actors. All subjects are without mustaches, eyeglasses, beards, earrings, visible make-up, wore grey T-shirts, and sat at a distance approximately three meters from the camera when the photo was taken.

The photo was taken from five different angles. However, in this research, a total of 980 images that consists of only the frontal face images were used as we aimed to detect landmarks of the frontal face. The fourth row of Fig. 4 shows examples of expressions from the KDEF database.

4) NIMH-ChEFS:The National Institute of Mental Health Child Emotional Faces Picture Set (NIMH-ChEFS) [20] is a dataset of posed facial images of children between the age of 10 to 17. The dataset consists of 534 pictures, where 341 pictures are images of girls and the remaining 193 pictures are images of boys.

Overall, the dataset consists of five different stimulus sets, which include afraid, angry, happy, neutral, and sad expres-

sions taken from an averted and direct angle. A total sum of 266 images from the NIMH-ChEF database, which are taken from direct angles is used in this research. The third row of Fig. 4 shows examples of expressions from the NIMH-ChEFS database.

5) RAF: The Real-world Affective Faces (RAF) [21] is a database of non-posed[2] facial expressions expressing the six basic plus neutral expression. The age distribution of the database ranges from 0 to 70+ years from three different races; i) Caucasian ii) African-American, and iii) Asian.

In this research, a total of 12271 images in the training set of the RAF database are used. The second row of Fig. 4 shows examples of expressions from the RAF database.

6) PICS: The psychological image collection at Stirling (PICS) [22] is a collection of databases containing a collection of posed images useful for conducting experiments. The dataset has no image of any famous person. All pictures are set to be 1200x1200 pixels and are taken from four different viewpoints, which are the straight, quarter left, half left, and full left viewpoints.

All 2D images of the PICS dataset that were taken from straight viewpoints are used in this research. The last row of Fig. 4 shows examples of expressions from the PICS database. Obligue images were not used as these features are not present in other datasets. Here, we consider distributions where the same features are plausibly present in all images.

Fig. 4. Sample expressions from the CK+, RAF, NIMH-ChEF, KDEF (ID: AF01, AF06, AF11, AF02, BM34, AM31, AM29), JAFFE, and PICS databases. Note: The first row represents images from the CK+, the second row represents images from RAF database, the third row represents images from NIMH-ChEF, the fourth row represents images from KDEF, the fifth row represents images from JAFFE database, and the last row represents images from PICS database. Also, across all databases (except NIMH-ChEFS dataset), images from left to right represent anger, disgust, fear, happy, neutral, sad, and surprise expressions. It is also worth noting that these expressions are only approximate and does not represent the internal states of a person.

C. Experimental Design

The extracted features of the landmark-based method are fed to a machine learning classifier; the SVM algorithm to enable fast computation. The SVM algorithm used is set to run with a kernel type k = linear, tolerance for stopping criteria = $1e-3$, and probability estimates p = True.

To test the performance of the landmark-based method, we test on all the databases one after the other. For instance, when testing on the CK+ dataset, we train on five databases (KDEF, JAFFE, NIMH-ChEF, RAF, and PICS) and test on the CK+, when testing on the KDEF dataset, we train on the CK+, JAFFE, NIMH-ChEF, RAF, and the PICS database and perform the testing on the KDEF dataset, and so on.

The performance of the landmark-based method is compared with four state-of-the-art deep models; EfficientNetB0, InceptionV3, ResNet50 and the VGG19 model. These models are set up to run for 200 epochs, starting with an initial learning rate of 0.001. The learning rate is scheduled to be reduced by 10% after 80, 120, 160, and by 5% after 180 epochs. The learning rate might also be reduced by monitoring the validation loss with the following parameter: factor f = 0.1, cooldown c = 0, patience p = 5, minimum learning rate min_lr = $0.5e-6$ to avoid over fitting the data.

We also compared the performance of the landmark-based method with Py-Feat and the Azure Face API to detect emotion from the six facial expression datasets used in this research. Py-Feat was configured to use "resmasknet" as emotion categorization model. The Azure Face API was configured to be used with the following parameters: recognitionModel = recognition_03, returnFaceAttributes = emotion, and as there are two detection models in the Azure Face API that can be used to detect faces in images [49], initially, the detectionModel parameter is set to detection_01, however, if the face was still not recognised using this model, the detectionModel is then set to detection_02 to detect the face. Neither of these models were able to detect certain faces of the RAF database and therefore certain images of the RAF database were not included in the experiment. Only the faces detected by either of these models were evaluated by the Azure Face API.

D. Experiment

The experimental results are presented in Table I and III. The performance of each method is evaluated using accuracy as the performance measure. Moreover, as the deep models used here are not deterministic, the results obtained by the deep models are presented with an upper and lower bound of a 95% confidence interval.

1) Comparison with state-of-the-art deep networks: As can be seen from the result obtained from Table I, the landmark-based method outperforms VGG19 and InceptionV3 in all cases when the test is performed on all the databases. The method also outperformed the ResNet model when the test is performed on all databases except for the PICS dataset.

[2]The non-posed pictures are sourced from the internet using different keywords in different languages and labeled by "emotion experts". Emotion experts are people that are trained to categorize emotions. These people categorize emotion based on the assumption that people smile when happy, frown when sad, and scowl when angry irrespective of their age, ethnicity, and gender.

TABLE I
RESULTS OBTAINED FROM CROSS-DATABASE EXPERIMENTS

Test on	Landmark-based	EfficientNetB0	InceptionV3	ResNet50	VGG19
CK+	51.84	49.49 ± 1.5	24.67 ± 1.8	41.35 ± 1.8	17.49 ± 0.9
JAFFE	53.52	41.60 ± 1.5	33.44 ± 1.5	43.47 ± 1.1	18.48 ± 1.2
KDEF	57.25	61.71 ± 1.1	35.55 ± 1.8	57.08 ± 1.8	17.99 ± 1.9
NIMH-ChEF	64.29	75.10 ± 0.9	52.97 ± 1.8	18.48 ± 0.5	37.16 ± 0.6
RAF	21.12	13.85 ± 0.9	16.55 ± 1.1	12.12 ± 0.9	18.99 ± 1.0
PICS	36.62	42.91 ± 0.8	32.50 ± 0.78	40.66 ± 0.8	13.60 ± 0.2

Moreover, we performed a statistical comparison with a two-sample t-test with Bonferroni correction to test if the result obtained by the landmark-based method is significantly different from any of the deep models at $\alpha = 0.0125$ (see Table II).

TABLE II
RESULTS FROM STATISTICAL COMPARISONS OF ACCURACIES

Test on	EfficientNetB0		InceptionV3		ResNet50		VGG19	
	t-value	p	t-value	p	t-value	p	t-value	p
CK+	-3.22	.003	-32.23	< .001	-12.36	< .001	-81.54	< .001
JAFFE	-17.55	< .001	-28.51	< .001	-19.42	< .001	-62.73	< .001
KDEF	+8.92	< .001	-25.72	< .001	-0.07	.944	-44.17	< .001
NIMH-ChEF	+25.63	< .001	-13.17	< .001	-195.92	< .001	-96.63	< .001
RAF	-17.15	< .001	-8.76	< .001	-21.27	< .001	-4.43	< .001
PICS	+18.08	< .001	-11.14	< .001	+10.938	< .001	-208.85	< .001

In Table II, the columns EfficientNetB0, InceptionV3, ResNet50 and VGG19 show the result of the comparisons between the landmark-based method and EfficientNetB0, InceptionV3, ResNet50 and the VGG19 model respectively.

The landmark-based method performed significantly better than the VGG19 and the InceptionV3 model in all cases. The method has also outperformed the ResNet50 model in four (on CK+, JAFFE, NIMH-ChEF, and RAF), and the EfficientNetB0 model in three (on CK+, JAFFE, and RAF) out of six different picture databases.

Overall, the landmark-based method has achieved a higher accuracy with an average accuracy of 47.44% across the six different databases compared with the deep models (except EfficientNetB0) that all achieved an average accuracy of less than 36%. Hence, these results suggest that the landmark-based method has a higher likelihood of achieving a better accuracy result than the deep models.

In addition, while it takes the landmark-based method approximately 30-45 mins to train, it takes approximately 3-7 hrs to train the deep learning models on the same machine.

2) Comparison with Emotion Categorization Tools: All the results obtained by the landmark-based method in Section IV-D1 are repeated here. Also, since both the landmark-based method, as well as an already trained Py-Feat and Azure Face API models are deterministic, only the achieved accuracies are presented as the standard deviation here is zero.

As can be seen, results obtained by the landmark-based method outperformed the results obtained by Py-Feat (from 2.53% up to 46.59%) and Azure Face (from 9.6% up to 33.46%) with a large margin in 4 out of 6 scenarios where the evaluation is made on JAFFE, KDEF, NIMH-ChEFS, and the PICS dataset.

TABLE III
TABLE PRESENTING THE RESULTS OF COMPARISONS WITH EMOTION CATEGORIZATION TOOLS

Test on	Landmark-based	Azure Face	Py-Feat
CK+	51.84	53.14	59.62
JAFFE	53.52	26.29	25.82
KDEF	57.25	45.51	50.10
NIMH-ChEFS	64.29	30.83	17.70
RAF	21.12	25.53	30.14
PICS	36.62	27.00	34.09

Although the accuracy obtained by Azure Face is greater than the accuracy obtained by the landmark-based method on the CK+ and the RAF dataset, the difference is small ($< 1.4\%$-CK+ and $< 4.42\%$-RAF dataset) when compared to the difference in the accuracy achieved by the landmark-based method on the remaining four datasets ($> 9.5\%$).

Not only did the landmark-based method outperform Py-Feat and the Azure Face in the overall average classification accuracy across all dataset, at the same time it ran faster than these techniques. It took an average of 9 mins - 4 hrs for Py-Feat to evaluate. However, as the Azure Resource Manager has a maximum limits of 1200 entries/hr, 1500 entries/hr, or 4194304 bytes depending on the subscription type [50], not all the images can be evaluated at the same time and therefore certain delay is expected. Here, 35 seconds delay after evaluating each image, resulted in an average evaluation time ranging from 2-119 hours on all the databases. Besides, the overhead training time of these models is not considered as they have already been trained.

V. DISCUSSION

Based on the results obtained in Section IV-D1 and IV-D2, we can see that the landmark-based method has achieved a higher classification accuracy (on average) when compared with both the Azure Face API, Py-Feat and to the state-of-the-art deep neural networks (except for the EfficientNetB0 model). Meanwhile we know that the Py-Feat toolbox has already been trained with the CK+ and JAFFE database. It is also plausible that the Azure Face API has been trained on these datasets as we do not know exactly which datasets the API has been trained on.

Although the difference in the accuracy obtained by the landmark-based method is not significant on two (KDEF and PICS) of the six datasets when compared to the ResNet50 model, the method achieves a higher accuracy result than the deep models in most cases. A possible explanation of why the landmark-based method performs better than the deap models could be because the deep models consider the color distribution within pixels in an image during the categorization task. However, since different databases might have different color patterns depending on where the image was captured (posed dataset) or sourced (non-posed dataset), it is not a surprise that the deep models achieve lower accuracy in a cross-database study compared to the landmark-based method that considers patterns of the face and level of expressivity of emotion when extracting features.

Major cloud providers such Microsoft and Google provide emotion categorization APIs. We have considered Google's emotion categorization API (i.e. the Vision AI) but it has only four emotion categories (i.e. anger, disgust, joy, and sorrow). These are notably different to the standard benchmark training as it uses only a subset of the emotions and therefore provides not a suitable comparison. Moreover, we have also considered other emotion categorization frameworks such as DeapFace [42]. However, in empirical studies, we found that the Py-Feat toolbox achieved a higher performance compared to DeapFace. For that reason, we only choose Py-Feat and the Azure Face API as our benchmark when comparing with other tools to gauge the performance of our algorithm.

Presenting a novel image to Py-Feat and the Azure Face API, the prediction returns in less than a second if we do not consider any overhead time of sending the request to the cloud, obtaining the prediction, and sending back the prediction to the user. However, the presented time in Section IV-D2 gives the order of magnitude of what a real-life user could expect.

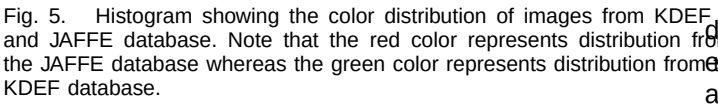

Fig. 5. Histogram showing the color distribution of images from KDEF and JAFFE database. Note that the red color represents distribution from the JAFFE database whereas the green color represents distribution from the KDEF database.

Considering that the deep models analyze the color distribution within pixels in an image, we plot a histogram showing the color distribution of randomly selected images from the JAFFE and KDEF database. As can be seen from Fig. 5, there is a large variation between the color distribution of images from the two databases.

Conversely, we find the difference of the extracted landmarks for the same images (from KDEF and JAFFE database) from Fig. 5 to see the variation in the extracted landmark coordinates. As seen from Fig. 6, the difference between the extracted landmarks of the images from these two databases (JAFFE and KDEF) weigh more around zero. Based on this observation, it is, therefore safe to say that the landmark-based method has a better chance of achieving a higher accuracy result in a cross-database study compared to the deep models as there is a high overlap in the extracted landmarks from the two different databases.

We know from our previous research [10] that the use of facial landmarks in emotion categorization leads to achieving a very high accuracy (>97%) on a single database. In contrast to that, we can see from Section IV-D that the achieved accuracy by the landmark-based method, although higher than the accuracy obtained by the deep models in most cases, is not very promising in the cross-database study. Therefore, we plot a histogram visualizing the extracted landmark features (x coordinate from all the six databases (see Fig. 7) to visualize the extracted landmarks.

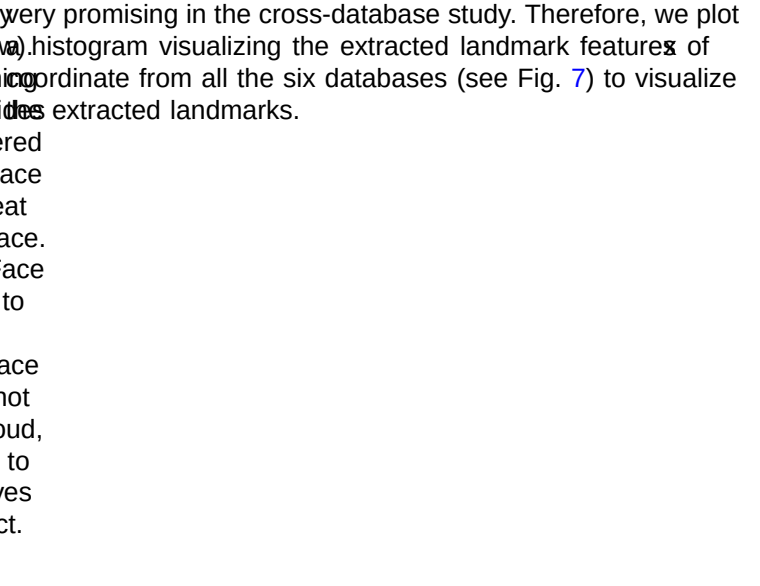

Fig. 6. Difference of extracted features from KDEF and JAFFE database

As can be seen, although there are overlaps, there is also variability in distribution of the extracted landmarks (x coordinate) from the different databases. Therefore, we can assume that this variation is what lead to the depreciation in the accuracy of the landmark-based method. Moreover, a one-way ANOVA was conducted to determine the significance of effect on the extracted x coordinates between the groups (i.e. the six databases) and a significance interaction was found [$F(5, 1292266) = 3757.22, p < .001$].

This is understandable as these databases are collected by different people, using a different set of instructions from an experimenter, and comprising of participants from different age groups, ethnicity, and gender. Also, certain subjects might have different (e.g. more oval or round) face shapes.

It is also worth noting that different deep models might have various resistance to different changes in the images. For instance, while the landmark-based method significantly outperformed the majority of the deep models used, the average accuracy achieved by EfficientNetB0 model across all the databases is equivalent with the average accuracy achieved by the landmark-based. Nevertheless, the landmark-based method (30-45 mins) is at least five times faster than the EfficientNetB0 (3-7 hrs) model.

This combination of findings provides support for the conceptual premise that deep learning algorithms, which are widely used for extracting features in emotion categorization are rather based on image classification than emotion categorization since emotions are not based on colors. End-to-end learning may focus on superficial features that do not extend to domains. Therefore, the findings have important implications for developing emotion categorization-based feature extraction methods that focus more on the pattern and level of expressiveness of an emotion rather than considering the color within the pixels in an image.

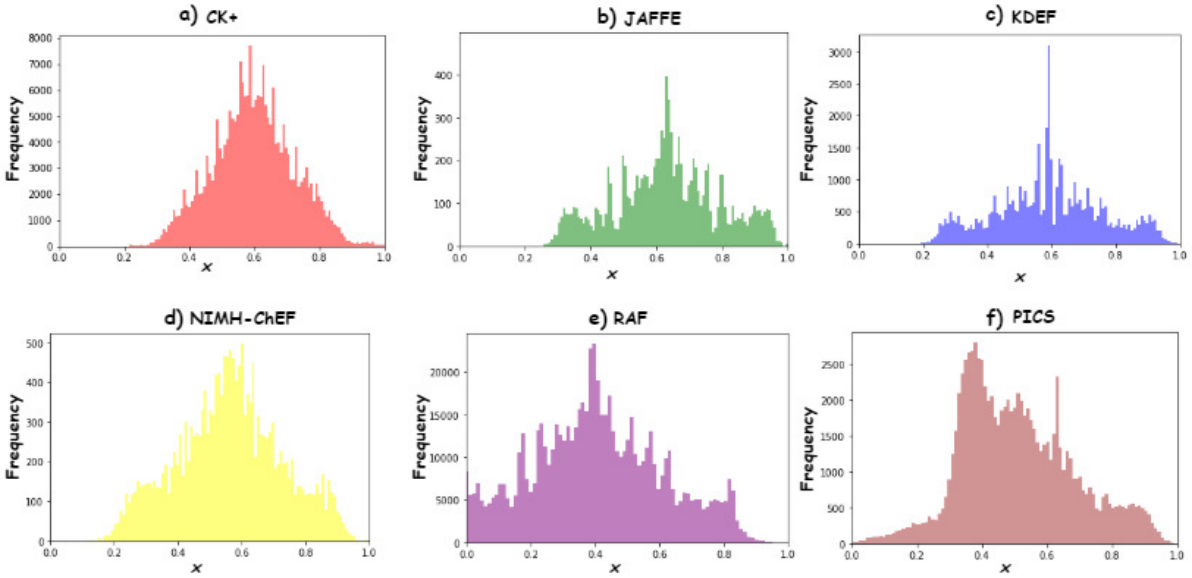

Fig. 7. Histogram showing the extracted features of x coordinate of the facial landmarks using the landmark-based method. Note that we have changed the y axis based on the maximum frequency per stimulus set for visualization purposes. Also, the x coordinates have been normalized.

VI. Conclusion

In this research, we analyze the robustness of a landmark-based technique in emotion categorization to an out-of-distribution attack. We compared the performance of the landmark-based method with two state-of-the-art deep models; ResNet50 and the VGG19 model, as well as the Azure Face API. The landmark-based method has demonstrated robustness in a cross-database study by achieving a higher accuracy than Azure Face, as well as performing significantly better than the deep models in most cases.

Although the landmark-based method outperformed the state-of-the-art deep models and the Azure Face API in most cases, the achieved accuracy is not very encouraging when tested on certain databases. For instance, the achieved accuracy is a little bit better than random guessing on the RAF database. This is an important issue to be addressed in future research. Hence, out-of-distribution emotion classification is a worthy area of study as it is difficult to categorize the emotion of people using only the superficial features from images. This also means that social robots may find it difficult to understand people' emotion from just their images. Further studies, which will augment the landmark-based method, e.g. temporal features, will need to be undertaken.

References

[1] F. Ren and Y. Bao, "A review on human-computer interaction and intelligent robots," *International Journal of Information Technology & Decision Making*, vol. 19, no. 01, p. 5–47, 2020.

[2] S. Li and W. Deng, "Deep facial expression recognition: A survey," *in IEEE Transactions on Affective Computing*, 2020.

[3] C. Nicholson-Smith, V. Mehrabi, S. F. Atashzar, and R. V. Patel, "A multi-functional lower- and upper-limb stroke rehabilitation robot," *in IEEE Transactions on Medical Robotics and Bionics*, vol. 2, no. 4, pp. 549–552, 2020.

[4] Y. Tang, X. Zhang, X. Hu, S. Wang, and H. Wang, "Facial expression recognition using frequency neural network," *in IEEE Transactions on Image Processing*, vol. 30, pp. 444–457, 2021.

[5] K. Khnlenz, B. Radig, C. Mayer, and S. Sosnowski, "Towards robotic facial mimicry: system development and evaluation," *in Proceeding International Symposium in Robot Human Interactive Communication*, 2011.

[6] S. Kim and H. Kim, "Deep explanation model for facial expression recognition through facial action coding unit," *IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 1–4, 2019.

[7] S. Zhang, X. Pan, Y. Cui, X. Zhao, and L. Liu, "Learning affective video features for facial expression recognition via hybrid deep learning," *in IEEE Access*, vol. 7, pp. 32 297–32 304, 2019.

[8] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, p. 966–979, 2012.

[9] C. Intahchomphoo and O. Gundersen, "Artificial intelligence and race: A systematic review," *Legal Information Management*, vol. 2, no. 20, pp. 74–84, 2020.

[10] H. A. Shehu, W. Browne, and H. Eisenbarth, "An adversarial attacks resistance-based approach to emotion recognition from images using facial landmarks," *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1307–1314, 2020.

[11] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *In ICLR, arXiv preprintarXiv:1409.1556*, pp. 770–778, 2015.

[15] J. H. Cheong, T. Xie, S. Byrne, and L. J. Chang, "Py-feat: Python facial expression analysis toolbox," *arXiv preprint arXiv:2104.03509*, 2021.

[16] A. Del Sole, "Introducing microsoft cognitive services," in *Microsoft Computer Vision APIs Distilled*. Springer, 2018, pp. 1–4.

[17] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, p. 94–101, August 2010.

[18] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces - kdef," *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institute*, 2017.