Multimedia Systems
https://doi.org/10.1007/s00530-023-01092-z

REGULAR PAPER

E-Cap Net: an efficient-capsule network for shallow and deepfakes forgery detection

Hafsa Ilyas[1] · **Ali Javed**[1] · **Khalid Mahmood Malik**[2] · **Aun Irtaza**[3]

Received: 4 August 2022 / Accepted: 9 April 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Deepfakes represent the generation of synthetic/fake images or videos using deep neural networks. As the techniques used for the generation of deepfakes are improving, the threats including social media disinformation, defamation, impersonation, and fraud are becoming more prevalent. The existing deepfakes detection models, including those that use convolution neural networks, do not generalize well when subjected to multiple deepfakes generation techniques and cross-corpora setting. Therefore, there is a need for the development of effective and efficient deepfakes detection methods. To explicitly model part-whole hierarchical relationships by using groups of neurons to encode visual entities and learn the relationships between real and fake artifacts, we propose a novel deep learning model efficient-capsule network (E-Cap Net) for classifying the facial images generated through different deepfakes generative techniques. More specifically, we introduce a low-cost max-feature-map (MFM) activation function in each primary capsule of our proposed E-Cap Net. The use of MFM activation enables our E-Cap Net to become light and robust as it suppresses the low activation neurons in each primary capsule. Performance of our approach is evaluated on two standard, largescale and diverse datasets i.e., Diverse Fake Face Dataset (DFFD) and FaceForensics++ (FF++), and also on the World Leaders Dataset (WLRD). Moreover, we also performed a cross-corpora evaluation to show the generalizability of our method for reliable deepfakes detection. The AUC of 99.99% on DFFD, 99.52% on FF++, and 98.31% on WLRD datasets indicate the effectiveness of our method for detecting the manipulated facial images generated via different deepfakes techniques.

Keywords Deepfakes · Efficient-capsule net · Diverse fake face dataset · FaceForensics++ · Synthetic face image detection

1 Introduction

Deepfakes refer to the generation of synthetic images or videos via deep neural networks. The term deepfake is a mixture of two words, "deep learning" and "fake" [1] and originated after a Reddit user named "deepfakes", who swapped celebrities' faces in pornographic videos using deep learning techniques [2]. Autoencoders and generative adversarial networks (GANs) are the deep learning models that are mostly used to generate deepfakes with the aim of creating more realistic images or videos [2]. Deep learning models based on the autoencoders use the autoencoder–decoder pairing structure where autoencoders extract the latent features from the face images and decoders are used to reconstruct the images [3]. But, in GAN-based deep learning techniques, two models (named generative model and discriminative model) are trained simultaneously. The generative model also known as the generator is used to generate fake images whereas the discriminative model known as the discriminator plays the role of detecting the fake images generated via the generator. The objective of a generator (G) is to capture the data distribution while the discriminator (D) estimates the probability of whether the incoming data is either from the training or the sample from G [4]. The availability of a variety of deepfakes apps (including ReFace, FaceApp, Face Swap Live, DeepFace Lab) has made it easy even for the less tech-savvy people to generate the deepfakes. FakeApp introduced in 2017 was the first attempt at deepfake creation. ZAO is another app that can swap the user faces onto

---

✉ Ali Javed
ali.javed@uettaxila.edu.pk

[1] Department of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan

[2] Department of Computer Science and Engineering, Oakland University, Rochester, MI 48309, USA

[3] Department of Computer Science, University of Engineering and Technology, Taxila 47050, Pakistan

Published online: 26 April 2023

movie star bodies and insert them into movies or TV clips [3]. Using the StyleGAN approach, the website [5] generates synthetic facial images with high-level realism. The commercially available deepfakes applications enable everyone to generate fake images and videos, which has increased concerns about circulating disinformation on social media, defamation, frauds, and hoaxes [2]. Besides the drawbacks, deepfakes also have productive and creative benefits including video dubbing of films, virtual try-on outfits, and education via reanimating the historical characters [1]. However, the excessive malicious usages of deepfakes suppress its positivity [3]. Therefore, reliable detection of deepfakes is very important and necessitates the development of tools that can effectively detect deepfakes images.

Deepfakes can be categorized as (1) face swap, (2) entire face synthesis, (3) face attribute manipulation, and (4) expression swapping [2].

- In face swapping, fake images or videos are created by swapping the face of a person with another person in the target image or a video retaining the background, expressions, and lighting [2]. The available models to create the swapped faces include FaceSwap [6], DeepFakes [7], and FaceShifter [8]. This type of manipulation can aid the film industry but can also be utilized for the wrong reasons such as financial fraud, hoaxes, etc. [3].
- Entire face synthesis includes the generation of realistic non-existing faces with high quality and is generated using the GANs. Recently, the StyleGAN approach is introduced to generate high-quality synthetic facial images that have a high level of realism. Such manipulation can be used for the creation of fake personas to spread disinformation on social media [1]. In the future, there exists a possibility that restoration methods such as GFP-GAN [9] can be used to suppress the appearance of forged content in GAN-generated images, thus, may make the detector job more difficult.
- In attribute manipulation, some face attributes (i.e., hair or skin color, gender, age, etc.) are modified. It is also known as face editing or retouching and can be used to try glasses, hairstyles, or makeup in a virtual environment [2].
- Expression swap involves the replacement of one person's facial expression with another in a video or image. An expression swap can be used to impersonate an identity as it allows one to animate the individual according to the attacker's desires [1].

In the last few years, many researchers introduced the methods and approaches that can detect fake facial images generated through deepfakes techniques. Marra et al. [10] presented an incremental learning model that can discriminate new GANs generated images without degrading the performance of previous ones. The disadvantage of this model [10] is that it performs well when various GAN models are available in the training phase. OC-FakeDect introduced in [11] was a one-class classification model based on variational autoencoder (VAE). The model was trained only on the real images, whereas tested on both the real and fake facial images. This approach [11] is only evaluated on FaceForensics++ (FF++) dataset and can be extended for images generated via GANs. Yuyang et al. [12] introduced a frequency in face forgery network (F^3-Net) that learned forgery clues via frequency-aware decomposition (FAD) and then extracted unusual frequency statistics among real and fake images through local frequency statistics (LFS). FAD and LFS features were then gradually fused to a module named as MixBlock. F^3-Net was evaluated on a challenging FF++ dataset and achieved an accuracy of 90% on the low-quality images. Most of the existing works focused on the detection of some specific manipulation techniques to determine the trustworthiness of facial images but failed to generalize their models on cross-corpora evaluation. Furthermore, most existing approaches for detecting deepfakes images are based on convolution neural networks (CNN) models and thus contain the drawbacks such as losing the features orientation and spatial information and not being equivariant, which means that CNNs cannot detect the images from different angles and rotated images if they are not trained on such images. Moreover, CNNs are unable to handle the Picasso problem (subject image with all the right components but not at the correct position) and often mislabeled such images.

The human brain analyzes the visual images through whole-part hierarchies such that it learns the features of the individual component and detects the orientation and relationship of the components in the whole subject image. To mimic the human brain's learning, Capsule Networks have been proposed that build the whole-part hierarchies using the neurons to encode the part and learn the relationship between the parts to detect the entire subject image, thus make the network interpretable and transparent. To address the aforementioned limitations of CNNs and existing deepfakes detection methods, we proposed a novel deep learning model efficient-capsule network (E-Cap Net) to efficiently and reliably detect the synthetic facial images generated through different deepfakes generative techniques. For shallow and deepfakes oriented synthetic facial images detection, the probability of object in the image and the orientation representing the parameters such as size, skin tone, object (i.e., nose, eyes, lips) orientation, and location in an image are important aspects that differentiate the fake face image from real one. In contrast to CNNs, our proposed E-Cap Net has the ability to learn these aspects for the classification of synthetic facial images. E-Cap Net can detect the rotated images taken from different viewpoints and also solves the Picasso problem. In our proposed model, we customized a capsule network and embedded a low-cost activation function max-feature-map (MFM) in its primary capsules. The embedded MFM activation function provides the compact representation of the features and enables our model to become light and computationally efficient. Moreover, our proposed model is capable of detecting multiple manipulation techniques including face swap, entire face synthesis, expression swapping, and face attribute manipulation. We also evaluated our approach for binary and multiclass classification problems. The major contributions of this work are:

1. We propose a robust Efficient-Capsule deep learning model containing the low-cost MFM activation function for accurate detection of shallow and deepfakes oriented synthetic images.
2. Our proposed model detects multiple types of deepfakes and is robust against varied deepfakes generation algorithms, different illumination conditions, ethnicity, age, images captured from different viewpoints, rotated images, and the Picasso problem.
3. We performed extensive experimentation on multiple datasets (covering multiple types of deepfakes) and also showed the efficacy of the proposed model against the existing state-of-the-art methods.
4. We also conducted the cross-corpora evaluation to show the generalization aptitude of the proposed E-Cap Net while detecting the shallow and deepfakes oriented synthetic images.

The remaining paper is organized as follows. In Sect. 2, we summarize the related work, while Sect. 3 presents our methodology for classifying the facial images either as real or fake. Experimental results are reported in Sect. 4. In Sect. 5, we provide the discussion. Finally, Sect. 6 presents the conclusion.

2 Related work

We reviewed the existing deepfake image detection techniques in Sect. 2.1, while Sect. 2.2 outlines the deepfake video detection. We also highlighted the limitations of existing deepfakes detection methods to present the knowledge gap in deepfakes detection.

2.1 Fake images detection

Initially, handcrafted features were commonly used to detect the discrepancies and artifacts in the fake images/video's synthesis process [2]. For example, Kim et al. [13] introduced a method that used local speed pattern (LSP) features to train the SVM classifier to detect fake and real facial images. Similarly, Xiaoqing et al. [14] utilized the universal steganalytic features in order to detect the images altered by various image processing operations. The extraction of meaningful, distinctive, and most appropriate handcrafted features is a difficult task as these features are constructed by domain experts and demand strong domain knowledge.

With the evolution of CNN, many researchers have applied deep learning techniques to extract the salient features automatically for image forensics. Bayar et al. [15] introduced convolution network architecture that detected different image manipulations and copy-editing operations without depending on the pre-selected features. In the same way, Rahmouni et al. [16] used a convolution network with a custom pooling layer to differentiate between the real and computer-generated visuals. The increasing use of CNNs has significantly enhanced the performance of deepfakes creation and detection, where models like autoencoders and GANs have made it possible to create photorealistic images and videos [17, 18]. In response to such photorealistic manipulated content, efforts have been made to develop effective methods to detect face forgery in images/videos [2]. Mo et al. [19] presented a CNN-based model that can identify progressive growing GAN (PGGAN) generated fake images and achieved an accuracy of 99.4% on the image size of 256 × 256. The accuracy of this model decreases to 96%, while reducing the image size to 128 × 128. Tariq et al. [20] introduced an ensemble ShallowNet classifier consisting of shallow layers to detect the fake face images created via the GAN. This model [20] was evaluated on different image sizes and performed well on small image resolution i.e., 64 × 64. These GAN detection models [19, 20] show good results when tested on images that are homogeneous to the training set images. In other words, the generalizability of these models is unknown. Nataraj et al. [21] presented a model that detected the manipulated images by extracting pixel co-occurrence matrices and then passed them to the CNN. To show the generalizability of the model [21], cross-validation was also performed. For this purpose, the cycleGAN images dataset (containing 35,302 images) was used to train the model, and then the trained model was tested on the StarGAN image dataset (containing 19,990 images) and vice versa. The lowest accuracy of 93.4% was attained with the model trained on the StarGAN image dataset as the classes were not uniformly distributed in the StarGAN dataset.

Besides the GAN-generated images dataset, researchers have also utilized other available datasets including FF++, DeepFake Detection, and Celeb-DF to evaluate their detection models. Zi et al. [22] presented an attention-based deepfake detection network ADDNet-2D for the detection of fake images. This model [22] consisted of ADD block followed by a 2D CNN network and a classification layer. Performance of this model was evaluated on 6 datasets including the DFD, DF-TIMIT (LQ, HQ), FF++ (LQ, HQ), and WildDeepfake. The highest accuracy of 99.82% was achieved on the FF++ HQ dataset, whereas achieved the lowest accuracy of 76.25% on the WildDeepfake dataset [22]. This model [22] is only evaluated on the DeepFakes subset of the FF++ dataset. AMTENnet introduced in [23] was a combination of AMTEN and CNN for detecting the manipulated facial images. AMTEN performed the preprocessing task to highlight discriminatory manipulated traces in the fake facial images. The manipulated traces were extracted by finding the difference between an input image and feature maps. Performance of this model [23] was evaluated on two datasets i.e., Hybrid Fake Face Dataset (HFF) and FF++. This work[23] performed the spatial filtering and lossy compression on the HFF dataset and then cross-validated those images but did not perform the cross-corpora evaluation on different facial manipulation techniques to evaluate the generalizability of their model. For the detection of forgery in facial images, Li et al. [24] introduced a detector that used a face X-ray (grayscale image) to find the discrepancies around the blending regions. The face X-ray detector is unable to perform well on the entire synthetic face as it relies on the presence of blending [24].

2.2 Fake videos detection

The detection methods used to identify fake images are not adequate to expose fake videos due to the frame data degradation and variable temporal characteristics between the set of frames [3]. Since digitally manipulated videos have temporal and intra-frame inconsistencies among the frames, Guera et al. [25] introduced a model that extracted frame features of a given video sequence using CNN and then passed the features to a long short-term memory (LSTM) network for analysis. Finally, a fully connected network was used to classify the video either as fake or real. For the evaluation of this model [25], 600 videos were gathered from different websites. Similarly, Sabir et al. [26] presented a pipeline consisting of two steps i.e., preprocessing and detection steps. Preprocessing step involved the detection, cropping, and alignment of faces in the frames while in the detection step, a recurrent convolution model (RCN) was used to identify the temporal artifacts between the set of frames. Along with the identification of temporal artifacts among video frames, researchers have also developed methods that detect the visual artifacts between the video frames to decide whether a given video sequence is manipulated or a real one [3]. Yang et al. [27] introduced a method that utilized 3D head poses to identify errors in a landmark location. Head poses were extracted using 68 facial landmarks. Difference between the estimated head poses was treated as a feature vector and passed to the SVM classifier for the detection of deepfakes. Matern et al. [28] presented a simple pipeline to exploit the artifacts that arise from the lack of global consistency, imprecise geometry, and illumination estimation. Missing reflections, eye color differences, and missing details in mouth and eye areas were used to detect the manipulated videos. Facial landmarks features were used with the logistic regression and neural network. Using this pipeline [28], this work detected the deepfakes, face2face manipulations, and synthetic faces and achieved the AUC values of up to 86.6%. The shortcoming of this pipeline is that it requires the images to have some specific prerequisites such as visible teeth and open eyes. The overview of the related work for fake images and videos detection is presented in Table 1.

2.3 Limitations of existing models

- Existing approaches are often evaluated on datasets with limited manipulation types, for instance, the FF++ dataset is limited to two fake types: expression and identity swap. Similarly, the Celeb-DF dataset only contains the identity swap fake images.
- Most existing approaches are based on CNNs and have some limitations including viewpoint variance problems and not being able to overcome the Picasso problem [29]. The reason is the use of Maxpooling layer for conveying the information from one layer to another. Therefore, the use of Maxpooling results in the loss of pose-aware and spatial information, thus hinders them to discover more about the image.
- Most of the work on the detection of synthetic facial images does not study the generalization capability of the models. So largely, existing detection methods fail to generalize well on cross-corpora evaluation which is an important requirement while developing a synthetic facial image detection considering the availability of multiple datasets and other repositories available in cyberspace.

3 Proposed method

This section presents the architectural details of our proposed deep learning model E-Cap Net. As an alternative to CNNs, Hinton et al. [30] first introduced the Capsule Network which is viewpoint invariant and identifies the whole entity via identifying its parts first. Capsule Network builds the whole-part hierarchies, represents the subject image as parts, and captures the relationships between the parts, thus making it more robust to the viewpoint variations of the input image. Capsule Network consists of low-level (primary) and high-level (output) capsules. The primary capsules in the network encode the information about the pose, scale, orientation, and other properties of the parts

Table 1 Overview of related work

References	Model/classifier	Dataset	Limitations
Fake images detection			
Mo et al. [19]	CNN	PGGAN	Poor results when reducing the image size
Tariq et al. [20]	ShallowNet classifier	CelebA PGGAN	Generalizability of model is unknown
Nataraj et al. [21]	Co-occurrence matrices + CNN	CycleGAN StarGAN	Performance degrades on jpeg compressed images
Zi et al. [22]	CNN	DFD DF-TIMIT (LQ, HQ) FF++ (LQ, HQ) WildDeepfake	Poor performance on WildDeepfake dataset Only consider the DeepFakes subset of FF++ dataset
AMTENnet [23]	AMTEN + CNN	HFF FF++	Not perform the cross-corpora evaluation on different facial manipulation techniques
Face X-ray [24]	CNN	FF++	Unable to perform well on entire synthetic faces
Fake videos detection			
Guera et al. [25]	CNN + LSTM	Private	Not robust against manipulated videos unseen during training
Sabir et al. [26]	CNN + RNN	FF++	Reported results only for static images
Yang et al. [27]	Landmarks + SVM	UADFV DARPA MediFor GAN image/video challenge	Performance degraded in case of blurry images
Matern et al. [28]	MLP + logreg	FF++	Applicable to the images having specific prerequisites e.g., open eyes, visible teeth etc.

in the subject image while the output capsules contain the information about the prediction. The output vector of the low-level capsules is routed to the appropriate high-level capsule through dynamic routing. The capsules in Capsule Network output a vector with the length representing the probability of object in the image and the orientation representing the parameters such as size, object (i.e., nose, eyes, lips) orientation, and location in an image. Therefore, unlike CNNs, there is no loss of orientation and spatial information in Capsule Network, as along with the feature detection, it also detects the orientation of features, texture, and color. Therefore, keeping in view the benefits of Capsule Network over the CNNs, we present a customized Capsule Network for the detection of manipulated facial images generated through different deepfakes techniques.

3.1 Architecture details

In our proposed E-Cap Net, after resizing the input image, the features are extracted from the resized image by utilizing a pre-trained VGG19 model. The extracted features are passed to the primary capsules in the customized Capsule Network and the outcome of the primary capsules is then passed to the output capsules through dynamic routing. Finally, the end results are calculated by computing the mean of activations of the output capsules. We customize Capsule Network via embedding the MFM activation function in each primary capsule, for the classification of shallow and deepfakes oriented facial images. The detailed architecture of our proposed model is shown in Fig. 1. The whole pipeline includes the input image, custom VGG19 to extract the features, Capsule Network, and final output.

3.1.1 Custom VGG19

The size of the input image is set to 300 × 300. The input image is passed to the custom VGG19 for extracting the features. The VGG19 is pre-trained using the ILSVRC database [31]. VGG19 has a total of 16 convolution layers that are used for feature extraction and 3 fully connected layers used for classification. The feature extraction layers are divided into five groups each followed by the max-pooling layer. We used the VGG19 to the third max-pooling layer for feature extraction with the hypothesis that lower level layers can preserve more information about the image. We used only the first eight convolutions layers of VGG19 for feature extraction. The benefit of using the custom pre-trained VGG19 network is that it aids in moderating the problem of overfitting. The summary of the used custom VGG19 model is shown in Table 2.

3.1.2 Efficient-capsule network

After the feature extraction, the extracted features are fed to the Capsule Network for the classification task. Our Capsule Network is comprised of primary and output capsules. It has

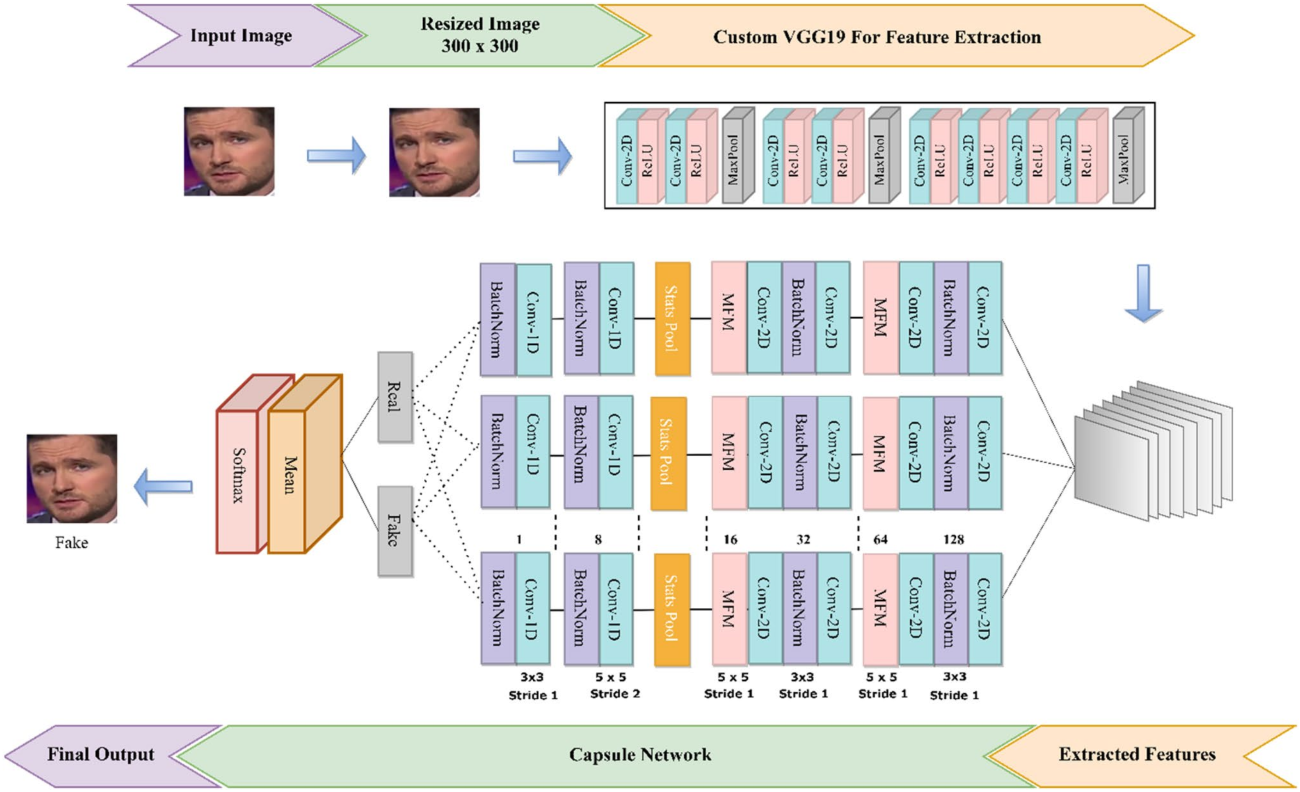
Input Image
Resized Image 300 x 300
Custom VGG19 For Feature Extraction
Conv-2D
ReLU
MaxPool
Stats Pool
MFM
BatchNorm
Conv-1D
Real
Fake
Softmax
Mean
Fake
1
8
16
32
64
128
3x3 Stride 1
5 x 5 Stride 2
5 x 5 Stride 1
3x3 Stride 1
5 x 5 Stride 1
3x3 Stride 1
Final Output
Capsule Network
Extracted Features

Fig. 1 Detailed architecture of proposed model

Table 2 Summary of custom VGG19

Layer (type)	Output shape	Param #
Conv2d—1	[− 1, 64, 224, 224]	1792
ReLU—2	[− 1, 64, 224, 224]	0
Conv2d—3	[− 1, 64, 224, 224]	36,928
ReLU—4	[− 1, 64, 224, 224]	0
MaxPool2d—5	[− 1, 64, 112, 112]	0
Conv2d—6	[− 1, 128, 112, 112]	73,856
ReLU—7	[− 1, 128, 112, 112]	0
Conv2d—8	[− 1, 128, 112, 112]	147,584
ReLU—9	[− 1, 128, 112, 112]	0
MaxPool2d—10	[− 1, 128, 56, 56]	0
Conv2d—11	[− 1, 256, 56, 56]	295,168
ReLU—12	[− 1, 256, 56, 56]	0
Conv2d—13	[− 1, 256, 56, 56]	590,080
ReLU—14	[− 1, 256, 56, 56]	0
Conv2d—15	[− 1, 256, 56, 56]	590,080
ReLU—16	[− 1, 256, 56, 56]	0
Conv2d—17	[− 1, 256, 56, 56]	590,080
ReLU—18	[− 1, 256, 56, 56]	0
MaxPool2d—19	[− 1, 256, 28, 28]	0

Table 3 Summary of primary capsule

Layer (type)	Output shape	Param #
Conv2d—1	[− 1, 128, 224, 224]	295,040
BatchNorm2d—2	[− 1, 128, 224, 224]	256
Conv2d—3	[− 1, 128, 224, 224]	409,728
MFM—4	[− 1, 64, 224, 224]	0
Conv2d—5	[− 1, 32, 224,224]	18,464
BatchNorm2d—6	[− 1, 32, 224,224]	64
Conv2d—7	[− 1, 32, 224,224]	25,632
MFM—8	[− 1, 16, 224,224]	0
StatsNet—9	[− 1, 2, 16]	0
Conv1d—10	[− 1, 8, 8]	88
BatchNorm1d—11	[− 1, 8, 8]	16
Conv1d—12	[− 1, 1, 8]	25
BatchNorm1d—13	[− 1, 1, 8]	2
View—14	[− 1, 8]	0

ten primary capsules, each having the same architecture. The summary of the internal architecture of a primary capsule is shown in Table 3. Each primary capsule further has five parts. The first two parts of each primary capsule encompass the convolution layer (Conv2d), batch normalization layer (BatchNorm2d), Conv2d, and MFM. The third part comprises statistical pooling while the last two parts include the convolution layer (Conv1d) and batch normalization layer (BatchNorm1d). For the convolution layers, we set the kernel size equal to 3 and stride of 1 except for the convolution layer after the statistical pooling layer for which we use a stride of 2 and kernel size is set to 5. Whereas for the MFM activation function, kernel size is set to 5, and a stride of 1 is used. MFM represses the activations of the neurons and thus enables the model to become robust and light [32], which helps to develop a computationally efficient model. Statistical pooling layer enables the network to extract the statistical properties by calculating the mean and standard deviation of frame-level features, which further helps in distinguishing the real and manipulated facial images. For statistical pooling calculation, we computed the mean and standard deviation as follows:

$$\mu = \frac{1}{mn} \sum_{i=0}^{m} \sum_{j=0}^{n} F_{ij}, \tag{1}$$

$$\sigma = \sqrt{\frac{1}{mn-1} \sum_{i=0}^{m} \sum_{j=0}^{n} \left(F_{ij} - \mu\right)^2}, \tag{2}$$

where μ denotes the mean, σ indicates the standard deviation, $m \times n$ represents the filter size and F represents the filter array.

There are two output capsules namely fake and real, for binary classification, whereas for multiclass classification, number of capsules depends on the number of classes available for classification. The outcomes of the primary capsules are routed to the output capsules via a dynamic routing [33]. Dynamic routing computes the agreement between outcomes of primary capsules and routed the obtained results to the appropriate output capsule (real or fake). Then the agreement for output capsules (real or fake) is calculated and the strength of the agreement determines the certainty of the label. The label is more certain if the agreement is stronger for an output capsule. The final output probabilities are determined based on the activations of neurons within output capsules. Finally, the softmax layer is applied to the output capsule vector to calculate the predicted label.

3.1.3 MFM activation function

To improve the classification performance and make the model computationally efficient, we implemented an activation function called MFM in each primary capsule of our E-Cap Net, instead of the traditional activation function (i.e., ReLU, Tanh). MFM is a variant of the Maxout activation function and delivers competitive feature maps rather than approximating convex activation from various feature maps. MFM has a sparse gradient and compact representation, thus allowing the model to become lighter. The sparse gradient can speed up the model convergence whereas compact representation can reduce the data dimensionality. This activation function divides the input layer feature map into two neurons unit at random and then output the element-wise maximum between the two units, which could reduce the non-relevant part of the feature map and can eliminate the redundancy in feature representation. The structure of the MFM activation function is shown in Fig. 2.

For an input convolution layer $c^n \in R^{W \times H}$, where $n = \{1, 2, \ldots, 2N\}$, H is the height, and W represents the width of the feature map, the MFM can be calculated as:

$$f\left(c_{xy}^{k}\right) = \max\left(c_{xy}^{k}, c_{xy}^{k+N}\right), \tag{3}$$

where $1 \leq k \leq N$, $1 \leq x \leq W$, $1 \leq y \leq H$, and $2N$ denotes the channels of the input layer.

4 Experimental results

In this section, we introduced the datasets and discussed the measures used to evaluate the performance of our proposed approach. We have performed extensive experimentation on the standard and diverse datasets for the evaluation of our model. The details of the experiments and their results are also discussed in the subsequent sections.

4.1 Datasets

We evaluated the performance of the proposed model on the World Leader Dataset (WLRD) [34] and on two standard, largescale and diverse datasets that are FF++ [35] and the Diverse Fake Face Dataset (DFFD) [36]. The details of these datasets are presented in the subsequent sections.

4.1.1 *FaceForensics++ dataset*

FaceForensics++ dataset is one of the largest deepfakes datasets and comprises 1000 original videos. These

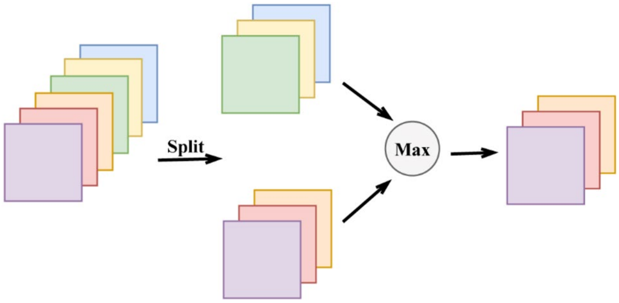

Fig. 2 Structure of MFM activation function

original videos are manipulated using different techniques including DeepFakes, FaceSwap, Face2Face, and NeuralTextures. The source of original videos is YouTube, and all the videos contain the frontal face of a person without any occlusions. The videos in FF++ dataset is available in three compression levels, i.e., raw (without compression), high quality (HQ, low compression), and low quality (LQ, heavy compression) [35].

To evaluate our model, we need an image dataset. For this purpose, we split the FF++ video dataset into training, testing, and validation sets. Training set contains 720 videos, while the testing and validation set comprises 140 videos each. Afterward, we extracted the faces from the available video's sequences (real and manipulated) to generate our image FF++ dataset. To generate our training set, we extract the first 100 frames of input video while for validation and testing, we extract only the first 10 frames. Shown in Fig. 3 are a few images from the FF++ dataset.

4.1.2 World leader dataset

WLRD comprises the videos having the FaceSwap manipulated images of different political leaders i.e., Obama, Hillary Clinton, Joe Bidden, Elizabeth Warren, and Bernie Sanders. Real videos are gathered from YouTube having only one person facing the camera. The comedic impersonator of the leaders is used to create the swapped faces. This dataset is highly imbalanced as it contains a very small number of fake videos as compared to real videos of each leader. The dataset is splitted into training, validation, and testing set. A few images from WLRD are shown in Fig. 4.

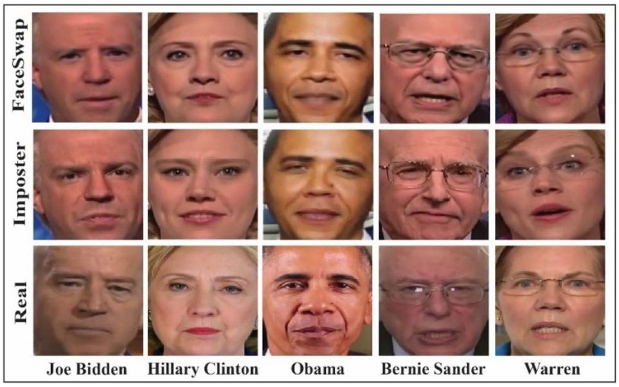

Fig. 4 World leader dataset

4.1.3 Diverse fake face dataset

DFFD as the name suggests comprises diverse types of fake faces which might be critical for the detection of face manipulations. DFFD comprises faces generated through StyleGAN, StarGAN, and PGGAN. DFFD also includes the real facial images of the FFHQ dataset. For real and each type of manipulated facial images, the dataset is splitted into 50% for training, 45% for testing, and 5% for validation. In DFFD, 47.7% images are of male subjects, while 52.3% of images are of female subjects and the age range of the subjects is 21–50 years [36]. Shown in Fig. 5 are a few images from the DFFD.

It is worth noticing that we performed our experiments on high-quality or low compression levels of the FF++ dataset. Moreover, for all the datasets (FF++, WLRD, and DFFD), training and validation images have never appeared in the test

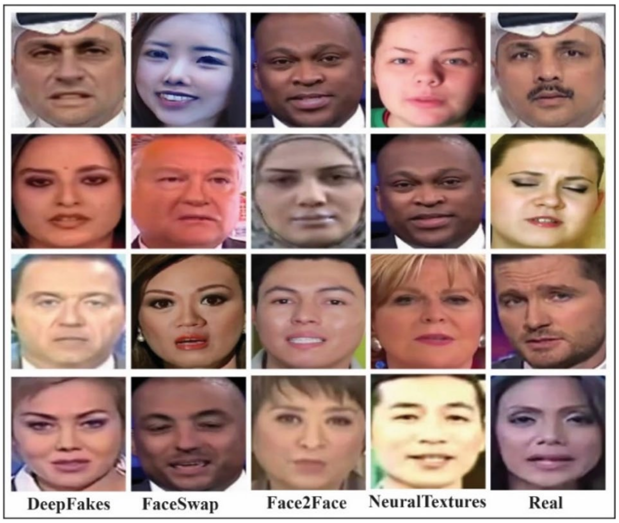

Fig. 3 FaceForensics++ dataset

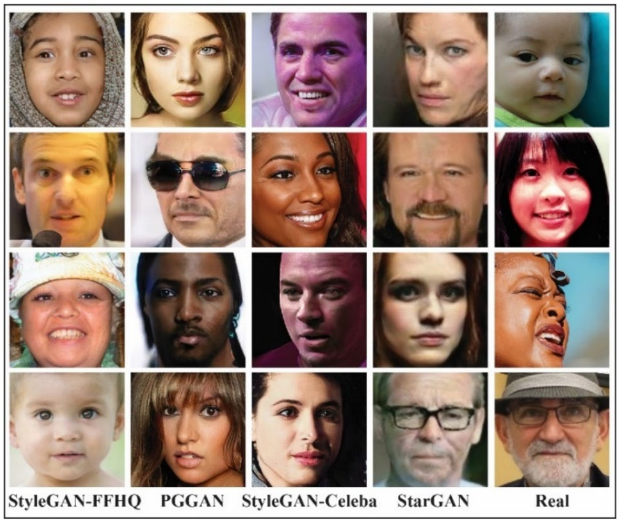

Fig. 5 Diverse fake face dataset

set. Thus, we tested our model on completely unseen images to show its effectiveness for manipulated facial image detection.

4.2 Implementation details

For the video dataset such as FF++ and WLRD, we used multi-task cascaded convolution neural network (MTCNN) [37] to extract the faces from the video frames. The extraction of faces from the video frames is a preprocessing step in the case of a video dataset. The proposed model implementation is based on PyTorch. All the images are resized to 300 × 300 resolution. The model is trained using an Adam optimizer with beta = 0.9 and learning rate = 0.0005. Other parameters are: batch size = 32, epochs = 25 and dropout = 0.05. For model implementation and execution, we used the high-performance computing machine with the following specifications: 4 NVIDIA Tesla V100 16G GPUs, 192 GB RAM, and 48 CPU Cores at 2.10 GHz.

4.3 Evaluation measures

To evaluate the performance of our proposed model, we use the following three evaluation metrics:

Accuracy represents the ratio of correctly predicted fake and real facial images to the total number of fake and real images in the test set. Accuracy is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{P + N}, \tag{4}$$

where *TP* represents the correctly predicted fake facial images and *TN* indicates the correctly detected real face images. *P* and *N* represent the total number of fake and real images, respectively.

Equal error rate (EER) represents the value at the point where the false acceptance rate (FAR) and false rejection rate (FRR) are equal. FRR represents the rate at which the model incorrectly classifies the fake images as real ones while the FAR refers to the rate at which the model incorrectly classifies the real facial images as manipulated ones. The lower value of EER represents the good detection performance of the model.

Area under curve (AUC) measures the classifier's ability to discriminate between the two classes (i.e., real and fake). It summarizes the classifier's performance by calculating the area under the receiver operating characteristic (ROC) curve. The higher AUC indicates better model performance in distinguishing between the two classes.

4.4 Performance evaluation of proposed method for real vs fake classification

To evaluate the performance of our method for the detection of real and fake/synthetic images, we designed an experiment to classify the real and fake images on FF++, DFFD, and WLRD datasets. In this experiment, we have a binary classification problem where we have two classes i.e., real and fake. The real class consists of pristine images while the fake class contains one type of manipulated images at a time. We split our datasets into three sets i.e., training, validation, and testing. For the training of our model, we used the training and validation sets. After that, the trained model is evaluated on the testing set to obtain the detection results. The results of this experiment in terms of AUC, EER, and accuracy on DFFD, FF++, and WLRD datasets are presented in Table 4.

Table 4 shows that our proposed model performs remarkably well on the DFFD dataset and achieved accuracy in excess of 99% for each type of GAN-generated fake facial images. Moreover, our model is able to detect the fake images generated through the StarGAN technique with 100% accuracy and AUC. These results indicate the effectiveness of the proposed model for accurately detecting the facial images having attribute manipulation. Overall, we can say that the proposed model can detect GAN-generated fake facial images with higher accuracy and less error rate. For the WLRD dataset, it can be seen that our proposed E-Cap Net accurately classifies the faceswap of the leaders with the AUC closer to 99% excluding Clinton for which the AUC is 93%. It can be observed from Fig. 4 that Clinton's imposter is closer to Hillary Clinton resulting in a more realistic swapped face, increasing the possibility of lower detection results compared to other leaders. We can observe that for the FF++ dataset, our proposed model detects the images generated through different deepfakes techniques with good accuracy and AUC. The AUC of DeepFakes and FaceSwap subset of FF++ is 98.61% and 99.51%, respectively. This remarkable performance on the faceswap manipulation

Table 4 Binary classification

Dataset	Face manipulation generation techniques	AUC	Accuracy	EER
DFFD	StyleGAN-FFHQ	99.96	99.59	0.24
	StyleGAN-Celeba	99.99	99.66	0.32
	PGGAN	99.92	99.99	0.08
	StarGAN	100	100	0.01
FF++	DeepFakes	98.61	97.17	2.50
	FaceSwap	99.51	98.68	1.50
	Face2Face	99.68	98.75	1.29
	NeuralTextures	95.14	91.61	8.25
WLRD	Hillary Clinton	93.05	92.29	16.08
	Joe Bidden	99.97	99.96	0.16
	Obama	98.87	98.28	2.16
	Bernie Sander	99.75	99.73	0.91
	Elizabeth Warren	99.91	98.51	0.45

indicates that our model has a strong capability of detecting swapped faces generated via different techniques. The detection accuracy of our proposed model is lowest on NeuralTextures images which is 91.61%. After a detailed investigation, we found that the NeuralTextures generate fake faces with very few semantic changes which are quite difficult to detect. This gives an indication that the detection of this type of manipulation is a challenging task.

Overall, it can be inferred from the results in Table 4, that the proposed E-Cap Net can accurately detect different types of manipulated images generated using different generative algorithms. This could be due to the fact that E-Cap Net captures the relative position and hierarchal relationship between different features in the facial image (such as eyes, nose, and mouth). Additionally, the MFM activation function can help the model to focus on the salient features in the given input image and reduce the impact of noisy or irrelevant information. Therefore, the proposed E-Cap Net can help to capture the compact and fine-grained details of input image allowing the network to better detect subtle differences between real and fake faces.

4.5 Performance evaluation of proposed method for multiclass classification

To examine the ability of our method for classifying multiple types of deepfakes, we designed an experiment to evaluate the performance of our model for multiclass classification problems on DFFD and FF++ datasets. In the case of FF++ multiclass classification, we have five classes named as DeepFakes, FaceSwap, Face2Face, NeuralTextures, and Real whereas, for DFFD multiclass classification, the classes are: StyleGAN-FFHQ, StyleGAN-Celeba, StarGAN, PGGAN and Real. We split both datasets into three sets i.e., training, validation, and testing. For the classification of the real and fake images, we trained our model using training and validation sets and then evaluated its performance on the testing set. For the FF++ dataset, the test set contains 1400 images, whereas for DFFD there are 9000 images in each class. The results of this experiment in terms of accuracy for each class on DFFD and FF++ datasets are presented in Table 5.

Table 5 Multiclass classification

Datasets	Classes	Accuracy
DFFD	Real	98.69
	StyleGAN-FFHQ	99.07
	StyleGAN-Celeba	97.69
	PGGAN	99.99
	StarGAN	100
FF++	Real	89.07
	DeepFakes	95.76
	FaceSwap	98.71
	Face2Face	96
	NeuralTextures	92.21

For the multiclass classification of DFFD, the proposed model achieves an overall detection accuracy above 99%, indicating its ability to detect GAN-generated images accurately. Table 5 shows that for multiclass classification of DFFD, the detection accuracy for StyleGAN-FFHQ and StyleGAN-Celeba falls to some extent as compared to the binary classification. The reason is that the images are generated through the same technique that is StyleGAN, the only difference between the two (StyleGAN-FFHQ and StyleGAN-Celeba) is the real images used to generate fake faces. So, there is a probability of misclassifying fake images generated via StyleGAN-FFHQ and StyleGAN-Celeba, in the case of multiclass classification. The overall detection accuracy for the FF++ dataset is 94% which indicates the good performance of our model in the case of multiclass classification. For the FF++ dataset, the accuracy of DeepFakes and Face2Face falls whereas the accuracy of FaceSwap and NeuralTextures increases slightly as compared to the binary classification. The reason is that in multiclass classification, there are more fake classes so the probability of misclassifying a fake image increases which have an impact on the detection accuracy.

4.6 Performance evaluation of proposed method on rotation attack

To check the effectiveness of our proposed E-Cap Net on the unseen rotated images, we designed an experiment, where we rotate the testing set images of different subsets of DFFD dataset at 11 different rotation configurations (30°, 45°, 90°, 120°, 135°, 180°, 210°, 225°, 270°, 300°, 315°). Then, the model trained on the respective subset of DFFD dataset is used to evaluate the rotated images. For this experiment, we also compared the performance of E-Cap Net with our existing CNN-based model namely InceptionResNet-BiLSTM (IR-BiLSTM) [38]. The results of the experiment in terms of average accuracy are shown in Table 6. It is important to note that the models are trained only on straight images, the rotated images are not included in the training. From Table 6, it can be observed that a decrease in the detection accuracy occurs for the rotated images when compared with the results of straight images. E-Cap Net classifies rotated images of different subsets of the DFFD dataset with an accuracy equal to or greater than 75%. This indicates the fairly good robustness of our model against the rotation attack. It is also inferred that E-Cap Net performs better than IR-BiLSTM on both rotated and straight images. E-Cap Net attained such reasonable detection results for the rotated images, because the proposed model builds the whole-part

Table 6 Performance of E-Cap Net on rotation attack

		Subsets of DFFD dataset			
		StyleGAN-FFHQ	StyleGAN-Celeba	PGGAN	StarGAN
E-Cap Net	Straight images	99.59	99.66	99.99	100
	Rotated images	75	77.91	80.45	78
IR-BiLSTM [38]	Straight images	90	97.86	97.86	99.52
	Rotated images	60.88	64.5	72.66	74.4

Table 7 Ablation study

Activation function in proposed model	Accuracy	Training time
Sigmoid	99.49	2 h 40 min
LeakyReLU	99.53	2 h 30 min
ReLU	99.84	3 h
MFM	**99.96**	**2 h**

The best results are in bold

Table 8 Comparison with existing methods using DFFD

Models	AUC	EER
Xception + Reg. [36]	99.64	2.23
VGG16 + MAM [36]	99.67	2.66
Representative forgery mining (RFM) [39]	99.96	–
E-Cap Net (proposed)	**99.99**	**0.41**

The best results are in bold

hierarchies, represents the subject image as parts, and captures the relationships between the parts. Which enables the model to become robust to the variations of the input image, not seen during the training time.

4.7 Ablation study

We conducted an ablation study to investigate the impact of various activation functions on the performance and efficiency of our proposed model in terms of accuracy and training time. We conducted this experiment to show that our proposed E-Cap Net model is more effective and computationally efficient than its variants. This experiment is performed on the StyleGAN-FFHQ subset of the DFFD dataset and experimental protocols are kept the same as mentioned in Sect. 4.4 for the DFFD dataset. We employed four activation functions i.e., ReLU, LeakyReLU, Sigmoid, and MFM in Capsule Network to compare the performance and computational cost. The results of this ablation study are presented in Table 7.

It can be observed from the results that the MFM activation function has the least training time and achieved the highest accuracy as compared to other activation functions i.e., ReLU, LeakyReLU, and Sigmoid. Our proposed E-Cap Net outperforms all its variants by achieving the accuracy of 99.96% and attained the accuracy gain of 0.12 from the second-best performing activation function i.e., ReLU. However, ReLU has the most computational cost since its training time is the longest of all. From the results in Table 7, it can be concluded that MFM is a low-cost activation function that makes the model light and more robust while detecting fake images. Thus, we can summarize that our proposed E-Cap Net with MFM activation function can accurately detect synthetic fake faces and is more efficient and robust compared to its other variants.

4.8 Comparison with existing state-of-the-art methods

To evaluate the performance and effectiveness of our proposed approach against existing state-of-the-art methods, we designed a two-stage experiment. In the first stage of this experiment, we compared the overall detection results of classification on two datasets (DFFD and FF++) as the existing methods only provide the overall classification results. To conduct the experiment, experimental settings are kept the same as mentioned in Sect. 4.5. For the DFFD dataset, we reported the overall AUC and EER of our proposed model for multiclass classification as done in the existing works. Likewise, for the FF++ dataset, we only reported the overall detection accuracy for comparing our model with the existing methods. In Table 8, we compared the results of our approach on DFFD with existing methods whereas Table 9 shows the comparison of results on the FF++ dataset.

From Table 8, it is noticeable that our proposed model achieved the best performance on DFFD than any other stated model. Thus, our proposed approach is able to detect entirely synthetic facial images with almost 100% AUC. In other words, our model outperforms in detecting the GAN-generated facial images with an accuracy of 99.92%. Moreover, the EER value of our model is lowest than the other stated methods, which also indicates a good detection performance. From Table 9, we can see that our proposed model achieved an accuracy of 94.51% which is the highest among

Table 9 Comparison with existing methods using FF++

Models	Accuracy
fCNN [40]	78.3
OC-FakeDect1 [11]	84.25
OC-FakeDect2 [11]	85.8
AMTENnet [23]	90.11
E-Cap Net (proposed)	**94.51**

The best results are in bold

the stated methods. As compared to the previous model [23], our approach increases the detection accuracy by 4%.

In the second stage of this experiment, we compared the results of our proposed E-Cap Net with [34] on the WLRD dataset. The experimental protocols are kept the same as mentioned in Sect. 4.4. Results in terms of AUC are demonstrated in Table 10. It can be noticed that our model outperforms in detecting swapped faces of Obama, Sanders, and Warren compared to [34] with the AUC gain of 3.87%, 3.75%, and 1.91%, respectively. However, for faceswap of Clinton, our AUC is slightly less than [34].

In general, our proposed method provides remarkable detection results on all datasets against the existing state-of-the-art methods, which shows its ability to detect different types of facial image manipulation. As we used two diverse datasets (DFFD and FF++) and a WLRD dataset for the evaluation of our proposed methodology, which are completely different from each other and contain the facial images generated through different deepfakes techniques. These datasets encompass fake facial images that cover categories of deepfakes (i.e., entire face synthesis, face swap, attribute manipulation, and expression swap). The good detection performance of our proposed model on all datasets reveals its ability to identify manipulated facial images generated through widely used deepfakes methods. Therefore, it is obvious that our model is not limited to the detection of specific deepfake technique but is able to detect various face manipulation techniques. This shows that our proposed model is generalizable and has the capability to detect manipulated facial images generated via several fake face generation techniques.

Table 10 Comparison with the existing method using WLRD

Models	AUC				
	Clinton	Joe Bidden	Obama	Sander	Warren
Agarwal et al. [34]	**95**	–	95	96	98
E-Cap Net (proposed)	93.05	**99.97**	**98.87**	**99.75**	**99.91**

The best results are in bold

4.9 Cross-corpora evaluation

To assess the generalizability of our proposed method, we also performed the cross-corpora evaluation. The main purpose of cross-corpora evaluation is to analyze the potential of the proposed method in real-world applications. We cross-validated the fake facial images generated through different deepfakes techniques. For this purpose, we designed two experiments, cross-set and cross-dataset. The details of the experiments are provided in the subsequent sections.

4.9.1 Cross-set

To evaluate the generalizability of our proposed model for subsets of the FF++ and DFFD dataset, we designed a cross-set experiment. This experiment is carried out in different phases for both datasets based on the combination of manipulated images subsets during training. There are four combinations of fake class for both datasets (DFFD and FF++). For FF++, the four combinations are: (1) DeepFakes + FaceSwap + Face2Face (DF + FS + F2F), (2) DeepFakes + FaceSwap + NeuralTextures (DF + FS + NT), (3) DeepFake + Face2Face + NeuralTextures (DF + F2F + NT), (4) FaceSwap + Face2Face + NeuralTextures (FS + F2F + NT). Whereas the four fake class combinations for DFFD dataset are: (1) StyleGAN-Celeba + StyleGAN-FFHQ + PGGAN (SGC + SGF + PGG), (2) StyleGAN-Celeba + StyleGAN-FFHQ + StarGAN (SGC + SGF + SG), (3) StyleGAN-Celeba + PGGAN + StarGAN (SGC + PGG + SG), (4) StyleGAN-FFHQ + PGGAN + StarGAN (SGF + PGG + SG). We trained the model on real and fake images where the fake class contains images from three subsets. The trained model is then evaluated on the remaining unseen subset. For instance, considering the DFFD dataset, in the first phase, we trained the model on real and fake images where the fake class contains three types of manipulated images (i.e., StyleGAN-Celeba, StyleGAN-FFHQ, and PGGAN). After that, an unseen subset i.e., StarGAN is used to evaluate the trained model and so on. The results of cross-set experiments for the DFFD dataset are shown in Table 11. Likewise, for the FF++ dataset, during the first phase, we trained the model on real and fake classes (containing fake images of FaceSwap, Face2Face, and NeuralTextures) and

Table 11 Cross-set evaluation on DFFD dataset

Training	Testing	Results	
		Accuracy	AUC
SGC + SGF + PGG	SG	96.62	99.31
SGC + SGF + SG	PGG	99.24	99.94
SGF + PGG + SG	SGC	99.63	99.96
SGC + PGG + SG	SGF	51.70	83.43

Table 12 Cross-set evaluation on FF++ dataset

Training	Testing	Results	
		Accuracy	AUC
FS + F2F + NT	DF	75.30	82.24
DF + FS + NT	F2F	63.46	70.93
DF + FS + F2F	NT	56.46	64.58
DF + F2F + NT	FS	48.50	45.98

then evaluated it on the unseen subset DeepFakes. The cross-set experimental results on the FF++ dataset are reported in Table 12.

As noted from Table 11, our proposed method outperforms with an AUC of 99% on the unseen subset of the DFFD dataset except for StyleGAN-FFHQ. Therefore, it is concluded that our proposed model trained on GAN-generated fake images outperforms in accurately detecting other unseen entire synthetic faces and attribute manipulated fake images. From Table 12, it can be observed that AUC drops when E-Cap Net is evaluated for totally unseen subsets of the FF++ dataset. The highest achieved AUC is 82% for detecting the DeepFakes subset as an unknown class while the lowest achieved AUC is 45.98% on the FaceSwap subset. The fact that only a small number of frames are manipulated in the FaceSwap subset could be the possible reason for lower AUC. All generative approaches utilized to create the fake faces in the FF++ dataset are completely distinct. For instance, Face2Face and FaceSwap are computer graphics-based methods for generating the manipulated facial images while DeepFakes and NeuralTextures are the deep learning-based approaches. Moreover, DeepFakes and FaceSwap include the face swap manipulation while Face2Face and NeuralTextures comprise expression swap manipulation. The unsatisfactory results in the case of FF++ dataset are attributed to the fact that the training and testing set in the cross-set examination use the synthetic faces generated via distinct and diverse fake face creation techniques. Thus, we can conclude that, in the case of cross-set experiments, our proposed model is capable of accurately detecting unseen synthetic facial images generated through other GAN-based techniques. This proves that our proposed approach has good generalization ability, especially for GAN-generated fake facial images.

4.9.2 Cross-dataset

To analyze the generalizability of our proposed E-Cap Net over distinct datasets, we performed a cross-dataset experiment using FF++ and WLRD datasets. The cross-dataset experiment has the following scenarios: (1) training on all subsets of FF++ dataset and testing on WLRD, (2) training on all subsets of FF++ dataset and testing on the Celeb-DF dataset [41], (3) training on all the subsets of WLRD and testing on FF++ dataset and (4) training on all the subsets of WLRD and testing on Celeb-DF dataset. The results are demonstrated in Table 13.

It can be observed from Table 13 that the proposed model trained on the FF++ dataset provides incredible results while detecting face swap manipulation of different leaders in WLRD. The highest achieved AUC is 99.39% and the lowest AUC is 76%, while detecting the swapped faces of different leaders. However, for the comedic imposter of different leaders, the results are slightly lower than face swap manipulation detection. The highest AUC of 87% is attained on the imposter of Joe Bidden. As the comedic imposter is a real person impersonating himself as the leader and not a synthetic content, this could be the possible reason for lower accuracy and AUC on the imposter subsets of different leaders. Likewise, E-Cap Net trained on the FF++ dataset when evaluated on the Celeb-DF dataset provides an accuracy of 83.65% and AUC of 67.94%. Additionally, testing accuracy of 69.38% and AUC of 57.24% are achieved on the Celeb-DF dataset for the model trained on the WLRD dataset. Celeb-DF dataset contains the high-quality realistic swapped faces with no color mismatch and decreased temporal flickering making the detection task more difficult. The dataset is highly imbalanced which can be the reason for low AUC value as compared to the accuracy. From Table 13, it is clear that the model trained on the WLRD dataset when evaluated on the FF++ dataset shows acceptable detection results except for the FaceSwap subset. The WLRD dataset contains the swapped faces generated through GAN-based algorithm. However, in the FF++ dataset, DeepFakes and FaceSwap are the subsets that contain swapped faces, the

Table 13 Cross-dataset evaluation

Training	Testing		Results	
			Accuracy	AUC
FF++	Clinton	Faceswap	92.32	99.14
		Imposter	55.60	67.27
	JB	Faceswap	97.71	99.39
		Imposter	84.59	87.06
	Obama	Faceswap	92.96	97.64
		Imposter	71.04	74.77
	Sander	Faceswap	73.03	74.09
		Imposter	74.03	84.36
	Warren	Faceswap	87.97	96.07
		Imposter	76.16	76.86
WLRD	DeepFakes		70.72	68.35
	FaceSwap		44.91	38.94
	Face2Face		51.32	52.48
	NeuralTextures		67.65	60.69

other two subsets (Face2Face and NeuralTextures) contain expression-swapped faces. The trained model shows good detection results for the DeepFakes subset, however, the poor detection result on the FaceSwap subset could be due to the reason that it involves the graphics-based method to generate the fake face. Moreover, the WLRD dataset is not diverse as it only contains the videos of five leaders, therefore, the model trained on this dataset is not robust enough to detect fake faces of diverse FF++ and Celeb-DF datasets. The remarkable detection results on the Faceswap subsets of WLRD and Celeb-DF dataset in the case of cross-dataset setting demonstrate the strong generalization ability of our proposed E-Cap Net especially when the fake face generation techniques are quite different from each other.

5 Discussion

The development of a robust model for detecting the shallow and deepfakes oriented synthetic facial images on diverse datasets and on cross-corpus settings is crucial to combat future viral disinformation campaigns. Moreover, lightweight deepfakes detectors are essential to deploy on resource constraint portable devices for real-time applications. Therefore, we introduced an E-Cap Net embedded with an MFM activation function that can detect fake facial images generated via diverse techniques (like DeepFakes, FaceSwap, Face2Face, NeuralTextures, StyleGAN, PGGAN, and StarGAN) with good accuracy. On the other hand, the MFM activation function provides sparse gradient and compact feature representation, resulting in a model that is light and converges quickly. The MFM activation function also makes the model efficient by reducing training and testing time as compared to other activation functions such as ReLU, LeakyReLU, and Sigmoid. We used only 25 epochs to train our model which is less than the number of epochs used in these methods [11, 40] indicating that less training time is required to train our proposed model E-Cap Net. Therefore, our model is more efficient than existing methods [11, 40].

Detailed analysis of the literature shows that the existing models are frequently evaluated on limited face manipulation types and also are not well generalized for the other manipulation types. For instance, the ADDNet-2D model in [22] was only evaluated on the DeepFakes subset of the FF++ dataset and does not provide better detection performance for other subsets included in the FF++ dataset. Moreover, the model in [23] is cross-validated for fake images altered using the image operations such as mean filtering and JPEG compression but not for fake images generated via diverse and different facial manipulation techniques. Our proposed model is evaluated on all the subsets of the FF++ dataset (i.e., DeepFakes, FaceSwap, Face2Face, and NeuralTextures) and also cross-validated for the detection of face images generated through different and diverse facial manipulation techniques (Sect. 4.9), thus solving the limitations of these models [22, 23]. It is important to mention that the proposed model demonstrates remarkable detection results for cross-set and cross-dataset evaluation. Especially, E-Cap Net trained on FF++ dataset provides remarkable detection performance in detecting the face swap manipulated images of WLRD and Celeb-DF datasets. However, the model trained on the WLRD dataset gives unsatisfactory results while detecting the fake images of FF++ and Celeb-DF datasets. This might be due to the fact that the WLRD dataset is small and not diverse as compared to FF++ and Celeb-DF in terms of manipulation techniques, age, and ethnicity.

Our proposed E-Cap Net is also robust towards the rotation attacks even though the model is not trained on such images (Sect. 4.6). The usage of the Capsule Network in our E-Cap Net model eliminates the problem of viewpoint variance, orientation and spatial information loss, and the Picasso problem that exists in the CNN. Capsule Network can handle viewpoint variance and Picasso problem since it models the hierarchical spatial relations between the different features of the image, rather than just detecting the features like CNNs. For instance, the network can recognize the relationship between facial features such as shape, position, and size of the eyes and nose, curvature of the mouth, etc. As the proposed model is based on Capsule Network, therefore, our E-Cap Net also learns the orientation and location of the parts (nose, eyes, lips, etc.) of real and fake faces. Thus, our model can capture the inconsistency in the orientation, position, and size of the components of the facial image, which enables the model to handle the images having all the mandatory parts but in the wrong place (Picasso problem). It is also important to note that our model provides good detection performance in the presence of varying illumination conditions, angled faces, and different ethnicities, ages, and gender.

Multiclass classification is another important aspect of deepfakes detection methods as it is more challenging than the binary classification problem. Besides the binary classification, our proposed E-Cap Net model also performed exceptionally well for multiclassification problem. Performance evaluation of our model for binary and multiclass classification (Tables 4, 5) shows remarkable results while detecting the GAN-generated images in both the binary and multiclass classification problems and thus overcomes the limitation of Face X-ray [24], which is unable to detect the entire synthetic faces. From Table 8, we can see that our proposed model brings a 2.25% decrease in EER over the model introduced in [36] for the DFFD dataset. On average, our proposed E-Cap Net model increases the detection accuracy by 10% on FF++ than the comparative methods shown in Table 9. Therefore, in general, we can say that our proposed model achieves remarkable results in detecting the manipulated facial images on all the datasets (FF++, DFFD, and WLRD), which proves that our model has the capability of detecting images generated through different facial manipulation techniques and is not limited towards the detection of specific fake face generation technique.

6 Conclusion

This paper has presented a novel and robust deep learning model E-Cap Net to detect the synthetic facial images generated through a variety of deepfakes algorithms. Our proposed E-Cap Net model has implemented MFM as an activation function which enabled it to become light and computationally efficient. Our model is capable of reliable detection of manipulated facial images in case of both binary and multiclass classification problems. We evaluated our proposed framework on DFFD, WLRD, and FF++ datasets and showed that the proposed model can effectively identify all types of manipulated facial images including face swap, expression swap, entire synthetic face and attribute manipulation. Extensive experimentation has also been accomplished to exhibit the generalizability of our model. We also demonstrate the effectiveness of our proposed model against rotation attacks. Experimental results indicated the great generalization aptitude of our proposed E-Cap Net for cross-set and cross-dataset evaluation. In the future, we plan to further improve the generalizability of the proposed model by making it more robust for all types of synthetic images. Additionally, we also intend to expand the model for the detection of deepfakes videos.

Acknowledgements This work was supported by the grant of the Punjab Higher Education Commission (PHEC) of Pakistan via Award no. (PHEC/ARA/PIRCA/20527/21), Michigan Translational Research and Commercialization (MTRAC) Advanced Computing Technologies (ACT) Grant Case number 292883, and NSF USA under Award no. 1815724. We would like to thank Prof. Hany Farid from the University of California Berkeley to provide us with their World Leaders Dataset for performance evaluation.

Author contributions Conceptualization—AJ, KMM; methodology—HI, AJ, KMM, AI; software—HI; validation—AJ, AI; formal analysis—AJ, KMM, HI, AI; investigation—AJ, KMM, AI; resources—AJ, KMM; data curation—HI, AJ; writing—original draft—HI, AJ; writing—review and editing—HI, AJ, KMM, AI; visualization—HI; supervision—AJ, KMM; project administration—AJ, KMM; funding acquisition—AJ, KMM.

Data availability statement FF++: FaceForensics++ dataset used during the current study is available at the following link https://github.com/ondyari/FaceForensics [35], DFFD: Diverse Fake Face Dataset is available at the following link http://cvlab.cse.msu.edu/dffd-dataset.html [36], whereas Celeb-DF dataset is available at the following link https://github.com/yuezunli/celeb-deepfakeforensics [41].

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR), **54**(1), 1–41 (2021)
2. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, **64**, 131–148 (2020)
3. Masood, M., Nawaz, M., Malik, K.M., Javed, A., Irtaza, A., Malik, H.: Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence, **53**, 3974–4026 (2022)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems (NIPS), **27**, 2672–2680 (2014)
5. Imagined by a GAN (generative adversarial network) StyleGAN2 (Dec 2019). "https://thispersondoesnotexist.com/", Accessed on July 10, 2022
6. *FaceSwap*. "https://github.com/MarekKowalski/FaceSwap/", Accessed on June 20, 2022
7. *DeepFakes*. "https://github.com/deepfakes/faceswap/", Accessed on June 20, 2022
8. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457, (2019)
9. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9168–9178
10. Marra, F., Saltori, C., Boato, G., Verdoliva, L.: Incremental learning for the detection and classification of gan-generated images. In: 2019 IEEE international workshop on information forensics and security (WIFS), 2019: IEEE, pp. 1–6
11. Khalid, H., Woo, S.S.: Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 656–657
12. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII, 2020, pp. 86–103. Springer
13. Kim, W., Suh, S., Han, J.J.: Face liveness detection from a single image via diffusion speed model. IEEE transactions on Image processing, **24**(8), 2456–2465 (2015)
14. Qiu, X., Li, H., Luo, W., Huang, J.: A universal image forensic strategy based on steganalytic model. In: Proceedings of the 2nd ACM Workshop on Information Hiding and Multimedia Security, 2014, pp. 165–170
15. Bayar, B., Stamm, M.C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, 2016, pp. 5–10
16. Rahmouni, N., Nozick, V., Yamagishi, J., Echizen, I.: Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE workshop on information forensics and security (WIFS), 2017: IEEE, pp. 1–6

17. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410
18. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in neural information processing systems, **33**, 12104–12114 (2020)
19. Mo, H., Chen, B., Luo, W.: Fake faces identification via convolutional neural network. In: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, 2018, pp. 43–47
20. Tariq, S., Lee, S., Kim, H., Shin, Y., Woo, S.S.: Detecting both machine and human created fake face images in the wild. In: Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, 2018, pp. 81–87
21. Nataraj, L., Mohammed, T.M., Manjunath, B.S., Chandrasekaran, S., Flenner, A., Bappy, J.H., Roy-Chowdhury, A.K.: Detecting GAN generated fake images using co-occurrence matrices. Electronic Imaging, **5**, 5321–5327 (2019)
22. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: Wilddeepfake: A challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2382–2390
23. Guo, Z., Yang, G., Chen, J., Sun, X.: Fake face detection via adaptive manipulation traces extraction network. Computer Vision and Image Understanding, **204**, 103170 (2021)
24. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5001–5010
25. Güera, D., Delp, E.J.: Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018: IEEE, pp. 1–6
26. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., Natarajan, P.: Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI), **3**(1), 80–87 (2019)
27. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: IEEE, pp. 8261–8265
28. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019: IEEE, pp. 83–92
29. Pawan, S., Rajan, J.: Capsule networks for image classification: A review. Neurocomputing, **509**, 102–120 (2022)
30. Hinton, G.E., Krizhevsky, A., Wang, S.D.: Transforming autoencoders. In: Artificial Neural Networks and Machine Learning—ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14–17, 2011, Proceedings, Part I 21, 2011, pp. 44–51. Springer
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, (2014)
32. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security, **13**(11), 2884–2896 (2018)
33. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. Advances in neural information processing systems, **30**, 3859–3869 (2017)
34. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting world leaders against deep fakes. In: CVPR Workshops, 2019, vol. 1, p. 38
35. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11
36. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, 2020, pp. 5781–5790
37. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, **23**(10), 1499–1503 (2016)
38. Ilyas, H., Irtaza, A., Javed, A., Malik, K.M.: Deepfakes examiner: An end-to-end deep learning model for deepfakes videos detection. In: 2022 16th International Conference on Open Source Systems and Technologies (ICOSST), 2022: IEEE, pp. 1–6
39. Wang, C., Deng, W.: Representative forgery mining for fake face detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14923–14932
40. Kohli, A., Gupta, A.: Detecting DeepFake, FaceSwap and Face2Face facial forgeries using frequency CNN. Multimedia Tools and Applications, **80**(12), 18461–18478 (2021)
41. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207–3216

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.