ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20231220161257

Preliminary communication
Received: 2023-12-20, Accepted: 2024-03-15

Advancing Image Forensic: Detecting Facial Manipulations via Meso_KNN

Hazem Issa, Bayan Zarnahji*

Abstract: Advancements in the fields of computer vision and deep learning have enabled the creation of highly realistic images, especially in generating human faces with an unprecedented level of realism. However, the misuse of these capabilities, such as in creating malicious content, has made image manipulation one of the most significant challenges in our daily lives. Therefore, it has become essential to develop innovative methodologies to distinguish between genuine and computer-generated multimedia, which continuously improves in terms of quality and realism. As a result, an effective model has been developed using deep learning techniques, relying on the deep neural network known as Meso Net and the K-nearest neighbors algorithm. This model, referred to as Meso_KNN, is presented in this paper. What distinguishes this model is its focus on important features in facial images that represent vital characteristics of facial manipulation. Additionally, it harnesses the capabilities of K-nearest neighbors for classification, achieving outstanding efficiency in detecting various types of facial manipulation. The model has been tested on a diverse set of facial images collected in the HFF dataset and has achieved an accuracy rate of up to 100 %. It stands as one of the current leading results in this field.

Keywords: deep learning; face image manipulation; machine learning; Meso Net

1 INTRODUCTION

Since the advent of digital visual media, there has always been a need to manipulate them for various purposes. Initially, manipulating media required expertise and consumed a significant amount of time and effort. With the advancement in computer technology and image processing software, modifying images, especially faces, has become much easier, enabling manipulations that were not possible in the past. Nowadays, anyone can create a face that looks realistic but is not from the real world, even without any prior experience in this field, thanks to the advancements in deep learning techniques and their capacity to produce high-resolution images using different versions of Variational AutoEncoders (VAEs) [1] and Generative Adversarial Networks (GAN) [2]. Despite being used in various applications such as gaming and filmmaking, their use in fabricating fake news, spreading it, and manipulating public opinion has become a tangible and significant threat to media information integrity. This could lead to harmful consequences. Deepfake, known for its ability to create and manipulate facial appearance (features, identity, and expressions) through deep methods, can be classified into the following 4 categories: 1) Entire Face Synthesis; 2) Identity Swap; 3) Expression Swap; 4) Attribute Manipulation [3]. The different manipulations of Deepfake present varying levels of risk. Entire Face Synthesis creates high-quality fake images of entirely imaginary individuals [4]. Identity swap involves replacing one person's face with another's; both techniques have the ability to change crucial personal information [5]. Manipulating facial features involves editing aspects like hair and skin colour, gender, age, and accessories, translating from image to image. Expression swap refers to creating images with specific facial expressions or replacing one person's facial expressions with another person's [6]. Techniques like identity swap replace the face of one person with another's, making actors appear in videos they never participated in; the face is replaced by exploiting comprehensive and adaptive facial information, with FaceSwap being one of the most famous applications. Some methods focus on whole-face synthesis, like PGGAN [7] and StyleGAN [8], aiming to produce and control highly detailed images up to 1024 × 1024, making it challenging to distinguish between real and fake. Expression Swap techniques propose generative models to create fake facial images without leaving any tangible traces, such as Glow [9] and GANination [10], using the effect of realistic images. Face2Face, relying on computer graphics (CG), animates facial expressions for the target video with facial expressions from the source face [11]. As for manipulating facial features based on GAN, applications like StarGAN [12] and AttGAN [13] can enhance feature editing. As shown in Fig. 1, in some examples of facial images, any user can now produce realistic synthetic faces that are challenging for humans to evaluate as real or fake. The impact that intentionally altered and maliciously used facial images have on people has made facial image manipulation detection an important problem in the field of Image Forensics.

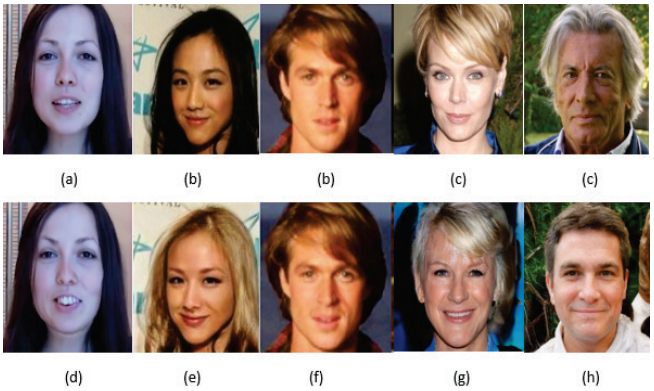

Figure 1 Example faces HFFD: Real images/frames from a) FaceForensics, b) CelebA, and c) CelebA-HQ datasets; Fake images generated by d) Face2Face, e) StarGAN, f) Glow, g) PGGAN and h) StyleGAN, respectively.

The purpose of this paper is to design an effective methodology for detecting various facial image manipulation techniques based on the Meso Net neural network and the K_nearest_neighbours' algorithm (Meso_KNN).

The Meso-4 model is an integrated model based on deep neural networks that perform well in classifying facial

images to detect facial manipulation [14]. Its main feature is the extraction of essential facial features for manipulation without any human intervention. It has a simple structure and is scalable and modifiable. The Meso model involves input layers, hidden layers, and an output layer. It is combined with K_nearest_neighbors in the output layer to improve classification accuracy. KNN is a fast and easy-to-use algorithm used for binary and multiclass classification and is usually employed to enhance system performance. Thus, the proposed hybrid model, Meso_KNN, consists of the Meso model and the KNN classifier connected in the Meso model's output layer. The following is a summary of the proposed model's contributions:

1) Proposing a composite model like Meso_KNN for detecting various facial image manipulation techniques based on Meso Net and the KNN classifier.
2) Achieving good results using the KNN classifier.
3) Comparing the results with other classification models on the same dataset, demonstrating that the final model performs highly in accuracy.

The paper's remaining sections are arranged as follows: presents related studies in Section 2, introduces the Meso_KNN model in Section 3, in Section 4 provides experimental results, conclusions, and future work in Section 5.

2 RELATED WORKS

Previous studies in the field of face manipulation detection have varied in their approaches, ranging from traditional techniques to modern artificial intelligence-based methods. These methods include the analysis of digital images and the use of neural networks to achieve more precise discrimination between real and manipulated images. In this paragraph, we will discuss some studies focusing on the techniques employed in face manipulation detection. Neural networks have been widely used to detect DeepFake manipulations. Li et al. (2020) utilised X-rays to identify the boundaries of manipulated faces; however, this method struggled to detect random noise and showed decreased performance with low-resolution images [15]. Dang et al. (2020) worked on identifying manipulated facial regions by estimating the attention map of the image [16]. They successfully detected visual manipulations; however, estimating the attention map in an unsupervised manner posed challenges. Models based on trainable convolutional neural networks (CNN) have demonstrated their ability to classify and recognise images, differentiating between manipulated and real images [17]. Transfer learning techniques from deep models like ResNet50 and VGG16 were integrated with CNN to enhance the model's performance [18]. In order to extract features related to manipulation, Afchar et al. (2018) built a model named Meso Net that included convolutional layer components that were optimised [14]. This model was characterised by its flexibility, allowing it to be modified and integrated with other networks. Guarnera et al. (2020) utilised the Expectation-Maximization (EM) algorithm to extract local features representing manipulated image effects [19]. SVM, LDA, and KNN were among the basic classifiers that effectively classified these features. These studies mainly focused on the development of models and their ability to utilise a diverse set of images for training and improving the model's performance. Therefore, we present the Meso_KNN approach, based on Meso Net and the KNN algorithm, to improve the performance of detecting different face manipulation techniques under a variety of complex and varied conditions, since the most successful methods heavily rely on CNN models.

3 PROPOSED APPROACH

Our methodology for detecting various face manipulation techniques involves analysing data based on the Meso Net and Meso_KNN models.

3.1 Meso Net Framework

Techniques that rely on analysing data at the micro, meso, and macroscopic levels are vital tools in the process of detecting facial manipulation. These techniques allow researchers to examine minute aspects and structural changes that can occur at the cellular and tissue levels. Microscopic-level analysis is used to study subtle changes on the skin surface and cells, focusing on pixel-level details to identify any alterations, while analysis at the Macroscopic-level is used to explore the biological and molecular details of facial structure manipulations. On the other hand, mesoscopic-level analysis combines the ability to detect fine changes and structural analysis simultaneously. It can identify manipulations that occur at both cellular and molecular levels accurately and comprehensively. The Meso-4 model is an efficient and integrated deep learning (DL) model designed for detecting face video forgeries. It was proposed in 2018, and the term "Meso" refers to the mesoscopic level of analysis, meaning it operates at an intermediate level for feature extraction between pixel-level and high-level semantic analysis. The goal of using Meso Net is to efficiently detect manipulation, making it suitable for real-time applications and scenarios that require processing large amounts of data using relatively few layers and parameters for high model performance. Meso Net consists of four consecutive convolutional layers, each followed by a Max Pooling layer and a batch normalization layer, this is followed by two fully connected layers [14]. Each convolutional layer extracts features, also known as feature maps. These feature maps are utilised for pixel value prediction as follows in Eq. (1):

$$y_i = \sum_{i=1}^{n} (x_i * w_{i,j} + b_i), \tag{1}$$

where b_i represents the bias term for the j^{th} convolution kernel, y_i represents the feature map output by the convolutional layer, and $x_i * w_{i,j}$ represents the convolution between the ith channel of the input image x and the ith channel of the j^{th} convolution kernel in the convolution operation. The feature maps are then passed through the ReLU activation function, as given in Eq. (2):

$$r(y) = \max(y, 0). \tag{2}$$

Where it changes the negative pixel values in the feature maps to zero, providing linear activation for neurons, neurons are not activated simultaneously. In order to extract maximum values from the feature map for critical data feature analysis, the corrected feature maps are also passed through a Max Pooling layer. Finally, a Batch Normalisation layer is employed to speed up convergence and lessen network overfitting.

At last, the learned features are forwarded to the classification unit, consisting of two fully connected layers. In the first layer, the linking between deep features is learned, containing 16 neural cells and a dropout layer to reduce sample dropout and enhance network robustness. In this study, we developed the Meso Net model, adding a Leaky ReLU layer [20] after the first fully connected layer. It is used in model training to speed up the training process, being efficient, and safeguarding the system from dying out due to the ReLU problem given in Eq. (3) below:

$$r(y) = 1(y < 0)(\alpha y) + 1(y \geq 0)(y). \tag{3}$$

Where α is a relatively small constant. The last fully connected layer's neurons contain an activation function that activates the required neural cells; the softmax activation function, which is used in multi-class output layer classification problems [21], was used in this research, as given in Eq. (4) below:

$$P = \frac{e^{zi}}{\sum_{j=1}^{n} e^{zj}}. \tag{4}$$

Where e^{zi} is the exponential function to measure the input ray, e^{zj} is the exponential function to measure the output ray, and P is the probability value. The output of the FC layer is responsible for predicting the final classification, where the output is [1 × 1 × N], where N indicates the number of categories. In the iterative training process, the error function is employed to minimize the loss between the true labels of the data and the network output during the training stage to make them converge. Categorical cross-entropy loss, which is a standard cost function for classification problems, aiming to reduce the gap between the expected and actual distribution to increase accuracy, is selected as the primary cost function in this study [22]. Meso adjusts the weights repeatedly to achieve the best results, as given in Eq. (5):

$$E = -\sum_{i=1}^{n} y_i * \log(P_i). \tag{5}$$

Where y_i represents the actual distribution for sample i, P_i represents the expected distribution given by the model for sample i and E represents the loss function. In binary classification P = 2, this indicates a real face image and a manipulated face image, respectively. In multi-class classification, P denotes the number of classes that better align with each type of image manipulation, with each class representing a method of face manipulation.

Given that the fully connected layer has all its neurons connected, passing input values and their weights is necessary for model training and classification. Training occurs through an iterative process involving forward passes of fed data and backpropagation. We randomly initialize the parameters of convolutional layers, then update the convolutional filters' weights and fully connected layers at each iteration of backpropagation. In this research, we employed Adaptive moment estimation (Adam) optimizer [23] to improve this loss value. Adam is a dynamic learning rate optimization technique that incorporates both scaling and momentum. It proves to be efficient and less memory demanding when dealing with intricate tasks involving extensive data or parameters. The formula for updating weights in Adam, represented as w_t for parameter w at time step t, as given in Eq. (6):

$$w_t = w_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \in}. \tag{6}$$

Where $\hat{m}_t$ represents the first moment average of gradients at step t, $\hat{v}_t$ represents the second moment average of squared gradients (uncentered variance), $\in$ represents a small constant to avoid division by zero and numerical issues, and η is the step size for optimizing the cross-entropy loss value.

3.2 Meso_KNN Framework

Illustrated in Fig. 2, the Meso_KNN framework primarily comprises two primary parts: the Meso part and the KNN part. The details of each of these parts are explained below:

3.2.1 Meso Net

The Meso part as illustrated in Fig. 2 is the main component of the model. Its primary objective is to extract crucial features from the dataset. With the same components mentioned earlier, except for isolating the last layer after the weights have been tuned for layers and replacing it with a KNN classifier for classification.

3.2.2 *k*-Nearest Neighbors' Algorithm

It is a simple and effective algorithm in machine learning, primarily used for classification problems. This algorithm utilizes the entire dataset as the training set instead of splitting it into training and testing sets because KNN algorithm is working on a pre-graded data separation, is based on the concept of nearest neighbors, where new points are classified based on the classes of their neighboring points. This significantly enhances classification efficiency and training speed.

Advantages:

- No need for prior training.
- Capability to handle non-linear data.
- Detection of outliers.

The algorithm is based on the principle of calculating the distance between items and all other items, then selecting a certain number of these items (k) as neighbors. These items are the closest ones to the pixel units in the dataset,

considering the calculated distance [24]. Euclidean distance is commonly used for distance calculation, as given in Eq. (7):

$$d(p,q)=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}. \quad (7)$$

Where n represents the total number of samples in the dataset, and p, q represents the distance between the unknown test sample and the known training sample. This function is used to calculate the similarity and dissimilarity between samples. Therefore, both real and manipulated face images match the output.

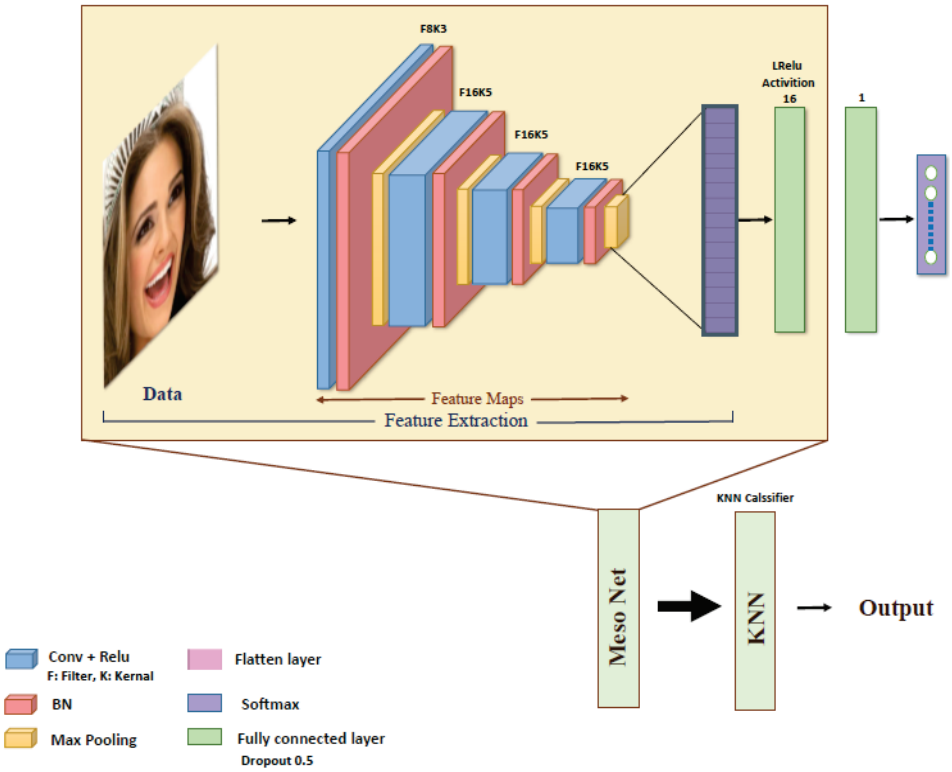

Figure 2 The Meso_KNN Proposed framework.

4 EXPERIMENTAL RESULTS

4.1 Experimental Settings

In this section, we will analyse the experimental results, evaluate the performance of our proposed model Meso_KNN and outperforms some other techniques.

4.1.1 Datasets

For the mentioned experiments, the Hybrid Fake Face Dataset (HFF) [25] dataset was used. It's a varied dataset for fake faces containing 8 types of face images. For real face images, it includes three types of images from three different open datasets. These types include low-resolution images from the CelebA dataset [26], high-resolution images from the CelebA-HQ dataset [7], and face video frames from the FaceForensics dataset [27], respectively. Thus, real face images are simulated under internet scenarios as realistically as possible. For fake face images, it includes PGGAN [7] and StyleGAN [8] for identity manipulation; it should be noted that both PGGAN and StyleGAN can generate no existing face images at a spatial resolution of 1024 × 1024; Face2Face [11] and Glow [9] for facial expression manipulation; and StarGAN [12] for transferring facial features, such as hair color and gender by multidomain image-to-image translation producing fake facial images. The HFF dataset is actually a large fake face dataset that consists of over 155K face images.

4.1.2 Implementation Details

The batch size was set to 32, and all face images in the dataset were resized to 64 × 64 for training the model. The detection accuracy on the test set was recorded after 100 epochs for model training.

4.1.3 Evaluation Criterion

The performance of these models is evaluated using Accuracy, Precision, Recall, and the F1-Score as the evaluation criteria. Accuracy in this case simply refers to how close the values of the model's predictions are to the actual (true vs. false) results. In other words, the number of times the model was able to accurately predict outcomes out of all the predictions it made. Eq. (8) shows the general formula used to calculate the prediction, where *TPR* represents the true predictions and *TOPR* represents the total predictions made by the model.

$$acc=\frac{TPR}{TOPR}. \quad (8)$$

On the other side, Precision (P) provides information about how consistent the obtained results are, even though they are not close to the actual values utilized by the target labelling. Eq. (9) illustrates the proportion of the identifications that were successfully made with respect to the actual ones. In Eq. (9), TP represents the number of true positives, and FP represents the number of false positives.

$$P = \frac{TP}{TP + FP}. \tag{9}$$

Recall (R) is the ratio of true positives correctly identified by the model to the total actual positives. Eq. (10) illustrates this ratio, where TP represents the number of true positives, and FN represents the number of false negatives. Recall intuitively captures the classifier's ability to find all positive samples [28].

$$R = \frac{TP}{TP + FN}. \tag{10}$$

The *F1-Score* works by taking both Precision and Recall into account, providing a balanced measure of a model's ability to predict both true positive and true negative instances. *F1-Score* can be interpreted as a harmonic mean of Precision and Recall. For the process of distinguishing between deep fake and real images, the *F1-Score* is the most suitable evaluation metric as both positive and negative classes are significant, and the relative contribution of Precision and Recall in *F1-Score* is better than equal weight. Eq. (11) illustrates how to calculate the *F1-Score* [28].

$$F1\text{-}Score = \frac{2(P \times R)}{P + R}. \tag{11}$$

4.2 Detection of Multiple Forgeries of Facial Image Manipulation

In our experiments, using the proposed model made it possible to simultaneously detect several facial image manipulations. The dataset contained different types of facial images that were randomly divided into two sub-groups for training (80%) and testing (20%), respectively.

In these experiments, the number of training images was approximately 124K, including five types of fake images and three types of real images with varying resolutions. For Meso_KNN model, we applied Adam optimizer for parameter estimation and the suggested default values for the moments, where the parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ were used in Meso Net training. In addition, the learning rate was set to 0.001 in order to reach the best categorical cross-entropy loss value [29] for the Meso Net. The results shown are for a Max Epoch value of 100, displaying the training progress of the Meso Net according to the illustrated graphs Figs. 3 and 4. The first graph shows the accuracy values, while the second one illustrates the loss values during the training period.

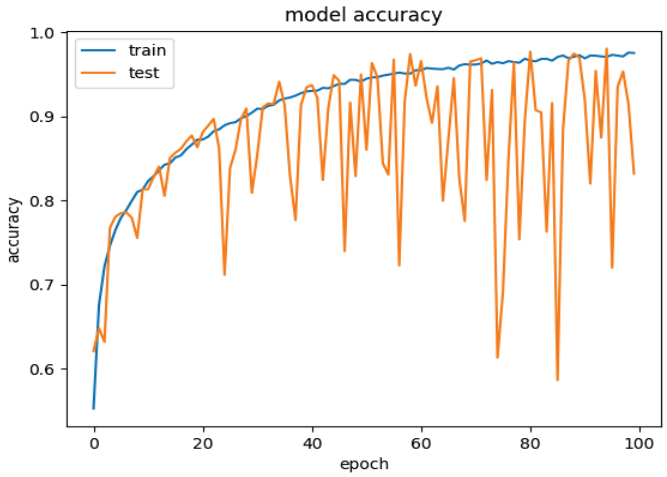

Figure 3 The variations in the accuracy of both the training and test samples during the training period of the Meso Net.

At the beginning of training, it's observed that the accuracy values start low but progressively increase at the end of the training period. Initially, the accuracy values for the test samples exceed those of the training samples, eventually converging towards the end of the training. In this study, the accuracy is reported at 98% for the training samples and 82% for the test samples for the Meso Net.

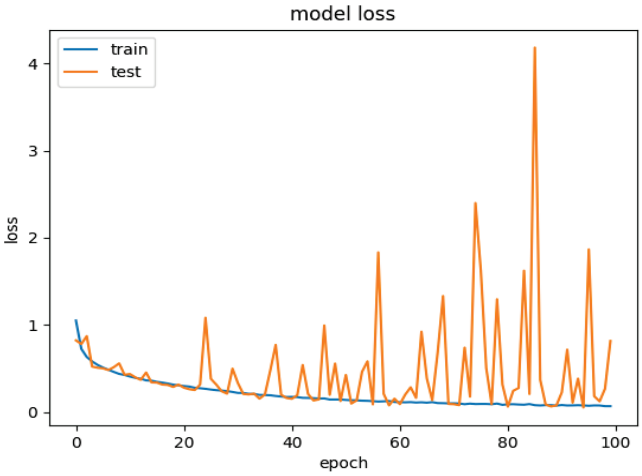

Figure 4 The variations in the loss of both training and test samples during the training period of the Meso Net.

At the beginning of training, it's observed that the loss values are relatively high, gradually diminishing at the end of the training period. Initially, the loss values of test samples are lower than those of training samples, gradually decreasing and fluctuating until they stabilize towards the end of training. In this study, the training samples achieve a loss value of 0.133, while the test samples achieve a loss value of 0.750 for the Meso Net.

A neighborhood size of k=10 and the Euclidean distance metric were adopted for KNN classifier in the Meso_KNN model. To evaluate the model's performance and its ability to separate and distinguish between categories, the following metrics were used: Precision, Recall, F1-score, and Accuracy.

Table 1 The rate of multiple classification identification achieved by the Meso Net model on HFFD calculated from the provided performance equations.

The class	Precision	Recall	F1-score
CelebA	52%	100%	68%
CelebA-HQ	96%	61%	75%
Youtube-Frame	90%	100%	94%
Glow	98%	31%	47%
StarGAN	96%	100%	98%
PGGAN	91%	71%	80%
StyleGAN	100%	95%	98%
Face2Face	100%	88%	93%

In Tabs. 1 and 2, the Precision and Recall rates for the proposed Meso_KNN model indicate that integrating KNN with the Meso architecture instead of the final fully connected classification layer had a positive impact on the model's behaviour. We evaluated the performance of the Meso Net and Meso_KNN models by creating confusion matrices in Tabs. 3 and 4, respectively. Their respective detection accuracies were 82 % and 100 %.

Table 2 The rate of multiple classification identification achieved by the Meso_KNN model on HFFD calculated from the provided performance equations.

The class	Precision	Recall	F1-score
CelebA	100%	100%	100%
CelebA-HQ	100%	100%	100%
Youtube -Frame	100%	100%	100%
Glow	100%	100%	100%
StarGAN	100%	100%	100%
PGGAN	100%	100%	100%
StyleGAN	100%	100%	100%
Face2Face	100%	100%	100%

Table 3 Confusion matrix to identify different types of manipulations using Meso Net. The asterisks "*" represent the value 0 %.

		Predicted class							
		CelebA	CelebA-HQ	Youtube -Frame	Glow	StarGAN	PGGAN	StyleGAN	Face2Face
The class	CelebA	82.18%	10.18%	0.08%	6.42%	*	0.08%	1.06%	*
	CelebA-HQ	1.85%	97.15%	*	*	0.8%	0.2%	*	*
	Youtube -Frame	0.2 %	*	99.96%	*	0.2 %	*	*	*
	Glow	72.66%	3.74%	*	23.12%	0.44%	0.04%	*	*
	StarGAN	0.1%	0.08%	0.2%	0.02%	99.76%	%0.02	*	*
	PGGAN	1.85%	96.75%	*	*	0.85%	0.4%	0.05%	0.1%
	StyleGAN	5.45%	7.85%	0.65%	*	0.05%	*	85.85%	0.15%
	Face2Face	*	*	32%	*	0.3%	*	*	67.7%

Table 4 Confusion matrix to identify different types of manipulations using Meso_KNN. The asterisks "*" represent the value 0 %.

		Predicted class							
		CelebA	CelebA-HQ	Youtube -Frame	Glow	StarGAN	PGGAN	StyleGAN	Face2Face
The class	CelebA	100 %	*	*	*	*	*	*	*
	CelebA-HQ	*	100 %	*	*	*	*	*	*
	Youtube -Frame	*	*	100 %	*	*	*	*	*
	Glow	*	*	*	100 %	*	*	*	*
	StarGAN	*	*	*	*	100 %	*	*	*
	PGGAN	*	*	*	*	*	100 %	*	*
	StyleGAN	*	*	*	*	*	*	100 %	*
	Face2Face	*	*	*	*	*	*	*	100 %

In Tab. 3, we can observe that false detection rates for Glow and CelebA are high. These two types of images share common characteristics, making it challenging for the model to distinguish, especially when resized to 64x64. Similarly, for PGGAN and CelebA-HQ, false detection rates are high for the same reason. On the other hand, the KNN classifier in the Meso_KNN model helped identify common features for every type, proving its effectiveness in identifying face manipulation.

4.3 Comparison of Performance

In this section, we compare the performance of the proposed model to that of other models by presenting experimental results on the given dataset.

Meso-4 [14]: It mainly exploits microscopic features of facial images to detect facial manipulation.

AMTEN [25]: AMTEN is used to extract facial manipulation artifacts for face manipulation detection.

Capsule [30]: It improves capsule networks to enable them to detect various types of counterfeits, such as the reuse of old data or computer-generated images and videos.

GRnet [31]: A guided residuals network utilises both spatial and residual information to better detect manipulated images.

Meso_KNN: The Meso_KNN model utilises both Meso data for analysis and feature extraction. Additionally, this model incorporates KNN classifier to enhance its efficiency and performance.

Table 5 Performance comparison for forensic models in multiple classification of different types of manipulation.

Methods	Accuracy, %
Meso-4	82
AMTEN	98.52
Capsule	96.75
GRnet	99.96
Meso_KNN (our)	100

Therefore, Tab. 5 illustrates a performance comparison between Meso_KNN and some recent works. We can notice that the proposed model shows higher accuracy compared to other models, achieving superior and top performance among current research efforts.

5 CONCLUSION AND FUTURE WORK

In this work, we proposed a novel methodology Meso_KNN for detecting manipulated images created by GANs. Meso_KNN is a robust framework based on the deep neural network Meso and the KNN classifier. Our proposed methodology is effective in image analysis and extracting important features for classifying fake and real images. Meso Net has proven its effectiveness in detecting images. Our model achieves a higher level of accuracy in detecting manipulated and real images in the HFF dataset. The results of our experiments have proven that the proposed method outperforms other techniques in terms of performance. Despite the high performance of the model on the HFF dataset, it is necessary to test the proposed model on other datasets. Although the system has been tested on RGB images, fake images using other color channels should be studied in the future, and this was not possible due to the limited resources available to us. This exploration is necessary to determine the process inputs for facial fake recognition.

6 REFERENCES

[1] Kingma, D. P. & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv*. https://arxiv.org/abs/1312.6114

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2672-2680.

[3] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148. https://doi.org/10.1016/j.inffus.2020.06.014

[4] Arjovsky, M., Chintala, S. & Bottou, L. (2017). Wasserstein GAN. *arXiv*. https://arxiv.org/abs/1701.07875

[5] Korshunova, I., Shi, W., Dambre, J. & Theis, L. (2017). Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 3677-3685. https://doi.org/10.1109/ICCV.2017.397

[6] Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W. & Wen, S. (2019). Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3673-3682. https://doi.org/10.1109/CVPR.2019.00379

[7] Karras, T., Aila, T., Laine, S. & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv* preprint *arXiv*: 1710.10196.

[8] Karras, T., Laine, S. & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401-4410. https://doi.org/10.1109/CVPR.2019.00453

[9] Kingma, D. P. & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 10215-10224.

[10] Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A. & Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, 818-833. https://doi.org/10.1007/978-3-030-01249-6_50

[11] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C. & Nießner, M. (2016). Face2face: Real-time Face Capture and Reenactment of RGB Videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 387-2395. https://doi.org/10.48550/arXiv.2007.14808

[12] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S. & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789-879.

[13] He, Z., Zuo, W., Kan, M., Shan, S. & Chen, X. (2019). AttGAN: Facial attribute editing by only changing what you want. *IEEE transactions on image processing, 28*(11), 5464-5478. https://doi.org/10.1109/TIP.2019.2916751

[14] Afchar, D., Nozick, V., Yamagishi, J. & Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. *In IEEE international workshop on information forensics and security (WIFS2018),* 1-7. https://doi.org/10.1109/WIFS.2018.8630761

[15] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F. & Guo, B. (2020). Face x-ray for more general face forgery detection. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5001-5010. https://doi.org/10.1109/CVPR42600.2020.00505

[16] Dang, H., Liu, F., Stehouwer, J., Liu, X. & Jain, A. K. (2020). On the detection of digital face manipulation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5781-5790. https://doi.org/10.1109/CVPR42600.2020.00582

[17] Tariq, S., Lee, S., Kim, H., Shin, Y. & Woo, S. S. (2019). Gan is a friend or foe? A framework to detect various fake face images. *In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 1296-1303. https://doi.org/10.1145/3297280.3297410

[18] Sharma, J., Sharma, S., Kumar, V., Hussein, H. S. & Alshazly, H. (2022). Deepfakes Classification of Faces Using Convolutional Neural Networks. *Traitement du Signal*, 39(3). https://doi.org/10.18280/ts.390330

[19] Guarnera, L., Giudice, O. & Battiato, S. (2020). Deepfake detection by analyzing convolutional traces. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 666-667. https://doi.org/10.1109/CVPRW50498.2020.00341

[20] Dubey, A. K. & Jain, V. (2019). Comparative Study of Convolution Neural Network's Relu and Leaky-Relu Activation Functions. In: Mishra, S., Sood, Y. & Tomar, A. (eds) Applications of Computing, Automation and Wireless Systems in Electrical Engineering. *Lecture Notes in Electrical Engineering, 553*. Springer, Singapore. https://doi.org/10.1007/978-981-13-6772-4_76

[21] Bishop Christopher, M. (2006). Pattern recognition and machine learning. *Information science and statistics*, New York: Springer.

[22] Gordon-Rodriguez, E., Loaiza-Ganem, G., Pleiss, G. & Cunningham, J. P. (2020). Uses and abuses of the cross-entropy loss: Case studies in modern deep learning. https://doi.org/10.48550/arXiv.2011.05231

[23] Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv* preprint *arXiv*: 1412.6980.

[24] Sutton, O. (2012). Introduction to *k* nearest neighbor classification and condensed nearest neighbor data reduction. *University lectures*, University of Leicester, 1.

[25] Guo, Z., Yang, G., Chen, J. & Sun, X. (2021). Fake face detection via adaptive manipulation traces extraction network. *Computer Vision and Image Understanding*, 204, 103170. https://doi.org/10.1016/j.cviu.2021.103170

[26] Liu, Z., Luo, P., Wang, X. & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*. https://doi.org/10.1109/ICCV.2015.425

[27] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. & Nießner, M. (2018). Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv* preprint *arXiv*: 1803.09179.

[28] Ferreira, A., Nowroozi, E. & Barni, M. (2021). VIPPrint: A large scale dataset of printed and scanned images for synthetic face images detection and source linking. *arXiv* preprint *arXiv*: 2102.06792.

[29] Dang, M. (2022). Efficient vision-based face manipulation identification framework based on deep learning. *Electronics*, *11*(22), 3773. https://doi.org/10.3390/electronics11223773

[30] Nguyen, H. H., Yamagishi, J. & Echizen, I. (2019). Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2019)*, Brighton, UK, 2307-2311. https://doi.org/10.1109/ICASSP.2019.8682602

[31] Guo, Z., Yang, G., Chen, J. & Sun, X. (2023). Exposing Deepfake Face Forgeries with Guided Residuals. *IEEE Transactions on Multimedia*, *25*, 8458-8470. https://doi.org/10.1109/TMM.2023.3237169

Authors' contacts:

Hazem Issa, PhD
Department of Computer Engineering,
College of Electrical and Electronic Engineering, University of Aleppo,
Aleppo, Syria
hazemisaa17@gmail.com

Bayan Zarnahji, Master's degree student
Department of Computer Engineering,
College of Electrical and Electronic Engineering, University of Aleppo,
Aleppo, Syria
bayanzernadjei@gmail.com