DEEPFAKE IMAGE & VIDEO CLASSIFICATION

[1]Omsai Koli, [2]Jay Tribhuvan, [3]Arjun Boraste
[123]Students
[123]Department of Computer Engineering
[123]MIT SCHOOL OF COMPUTING, LONI, PUNE
[1]jaitribhuvan83@gmail.com, [2]jaitribhuvan83@gmail.com, [3]arjunboraste9611@gmail.com

***Abstract*—Fake images, also known as "deepfakes," are a growing concern in today's digital age. These images are often created with the intent of benefiting one party and can be difficult to distinguish from real images.**
They are often disseminated through digital media and newspapers, and can spread misinformation or propaganda, which can have serious consequences if not detected and addressed.
To effectively detect image falsification in many image data, an architectural model that can process several pixels in the image is required, as well as a method that is effective and adjustable with traning data for practical use in daily life.
In this paper to detecting fake images usingVGG19 is a convolutional neural network (CNN) architecture that has been successful in a variety of image classification tasks. The proposed VGG19 is better model compared existing models it provides 96% accuracy.

I. Introduction

In recent years, the identification of deepfake photos has become an increasingly relevant problem due to the proliferation of the usage of deepfake technology, which allows for the creation of fake images that seem to be very realistic. These pictures may be used to a number of nefarious uses, such as the dissemination of false information, the assumption of another person's identity, and the production of sexually explicit content without their consent. As a consequence of this, recognizing and locating deepfake pictures is a significant challenge that calls for the use of sophisticated methods. There have been a number of previous studies on the identification of deepfake images, the vast majority of which include the use of deep learning techniques. Deep convolutional neural networks, also known as DCNNs, are a common method that may be used to determine if a picture is genuine or false by analysing the patterns that are included within it. One example of such a DCNN is known as VGG19, and it is a sort of model that has been used in several different studies to identify deepfake images. Image classification and object recognition are only two of the many applications for VGG19, which is a convolutional neural network that has been trained to identify patterns in pictures.

To use VGG19 for deepfake picture identification, a dataset consisting of genuine and fake images that have been appropriately annotated must be gathered. After the photos have been pre-processed to ensure that they are in a consistent format, the VGG19 model is trained on the dataset using a supervised learning technique. This step takes place after the images have been processed. The deep neural network is trained using actual pictures throughout the process of training, and the output layer is taught to predict "real" as the output of

the network. Another deep neural network is trained in the same way using the false photos, and this time, the output layer is taught to predict ”fake” as the output of the network. After it has been trained, the model may be used to the task of determining whether or not fresh photos are genuine, or phony based on the patterns it has learnt to recognize during training. The VGG19 model offers several benefits when it comes to the identification of deepfake images. To begin, it can recognize intricate patterns in the data thanks to its enormous capacity and vast number of parameters, both of which make this capability possible. In addition, the VGG19 model has already been pre-trained on a large dataset and is capable of being fine-tuned such that it excels at a particular job. This can make it more computationally efficient compared to training a model from scratch, as the model can start with a set of learned features, and then f ine-tune them for the task of deepfake image detection. This is because the model can start with a set of learned features, and then fine-tune them for the task. For the purpose of detecting deepfake images, in addition to the use of deep learning strategies such as VGG19, additional approaches have been developed.

The following are examples of some of these methods: Performing an analysis on the artifacts that may be seen in the picture. Deep fake photos, for instance, are known to sometimes display visual distortions like blurriness, which may be used to identify them as such. Using methods from the field of signal processing in order to identify shifts in the audio or video signal that are characteristic of deepfake pictures. For instance, the Face Forensics dataset has movies that have been modified in a number of ways using a range of different methods. obtaining information from the picture itself, such as through analysing its textures, contours, and lighting to determine what’s going on in the scene. The rest of the paper is organized as follows section-2 describes literature survey, proposed work was discussed in section-3, section-4 describes the experimental results and section-5 concludes paper.

II. Existing Work

There are a few existing systems related to our project field. After some research and analysis, we came across the methodology of the system and a few of its drawbacks. The below table gives us the gist about these existing systems:-

Existing Model / Work	Description (What it Does)	Limitation
CNN-Based Fake Image Detection	Uses basic convolution layers to learn features and classify images as real or fake.	Struggles to detect high-quality deepfakes; limited detail extraction.
DenseNet Model	Connects each layer to every other layer to improve feature reuse and gradient flow.	Requires higher computation and memory; training is slow.
Capsule Networks (CapsNet)	Learns spatial relationships between image parts, improving forgery detection accuracy.	Hard to train on large datasets and requires high processing power.
Bi-LSTM + CNN Hybrid Model	Extracts visual + sequential patterns for fake news / fake media detection.	Works better for text + video combination; less accurate for image-only deepfake detection.
AFIFN (Advanced Fake Image-Feature Network)	Uses dual-layer deep learning to analyze forged image features.	Performance drops when images are heavily compressed or blurred.

III. Motivation

In today's digital world, deepfake images are increasingly used to spread misinformation, manipulate public opinion, impersonate individuals, and even commit fraud.

These fake images are very realistic and difficult to recognize with the naked eye, which makes them a serious threat to privacy, security, and trust in online content.

Therefore, there is a growing need for an accurate and automated system that can detect whether an image is real or artificially generated. This motivated us to develop a deep learning-based approach using the VGG19 model, which can learn fine visual features and effectively distinguish real images from deepfakes.
By implementing this system, we aim to help reduce misuse of manipulated images and improve digital media authenticity.

IV. Objectives

1. To detect and classify images as real or fake using a deep learning-based approach.
2. To use the VGG19 model for extracting detailed visual features for accurate deepfake identification.
3. To train and evaluate the model using real and artificially generated image datasets.
4. To improve the accuracy, precision, recall, and F-score compared to existing methods.
5. To provide a system that helps prevent misinformation and identity misuse caused by deepfake images.

V. Objectives

1. The project focuses on fake image detection (deepfake faces), not video deepfake detection.
2. Only images are considered from datasets like Flickr (real) and StyleGAN (fake).
3. The system is applicable to digital media platforms, verification systems, and security applications.
4. The model can be further extended to real-time detection or combined with video processing.
5. The scope includes training, testing, and evaluating the model, but does not include deployment into a live application.

Chapter 2

VI. CONCEPTS AND METHODS

This project focuses on detecting **deepfake images**, which are artificially generated or manipulated visuals created using advanced deep learning techniques. Such fake images can be used to spread misinformation, impersonate individuals, or influence public opinion. Therefore, it is necessary to develop a reliable system to differentiate **real images from fakeones**.To achieve this, the project uses **VGG19**, a deep Convolutional Neural Network known for its strong ability to extract fine visual details. A dataset containing **real face images** (e.g., Flickr dataset) and **fake face images** (generated using StyleGAN) is collected. All images are **pre-processed** by resizing and normalizing them for consistency.

The **VGG19 model is fine-tuned** on this dataset so it can learn the distinguishing features between genuine and manipulated images. The data is divided into training, validation, and testing sets to improve learning and avoid overfitting. After training, the model classifies any new input image as **Real** or **Fake**. The results show high accuracy, precision, recall, and F-score, proving that VGG19 performs effectively in deepfake detection.

Chapter 3

VII. LITERATURE SURVEY

According to research Kaliyar et al. [1], For detecting false news, the suggested model (FNDNet) is a deep convolutional neural network that automatically learns discriminating characteristics through many hidden layers. It achieved state- of-the-art results with an accuracy of 98.36% on test data using various performance evaluation parameters such as Wilcoxon, false positive rate, true nega tive rate precision-recall F1 score, etc., demonstrating significant improvements over existing models used for detecting Fake News from social media platforms. The capacity to learn discriminative features in a single run and the absence of manual feature extraction are the key benefits of this method.

However, training on big datasets may be difficult, which can increase processing time and computational expenses. Goldani et al. [2] Capsule neural networks were presented as a technique for identifying disingenuous articles. We employed multiple embedding models based on the length of a specific news item and used varying degrees of n-grams as features in our suggested model. In order to properly interpret and categorize text, these models' ability to record links between com ponents of sentences is crucial. During the training phase, they also provide incremental uptraining, which facilitates rapid adaptation to newly introduced data points or characteristics.

Our proposed model was shown to outperform existing methods by 7.8% on the ISOT dataset and 3.1% on the LIAR validation set with a 1% improvement over the test set accuracy from the LIAR dataset compared to state- of-the-art techniques currently available in this field; however, there may be some limitations such as computational complexity due its deep learning architecture or potential bias if not trained properly using diverse datasets representing all types of content related topics accurately, Kumari and Ekbal [3] proposes a multimodal methodology for detecting false news that uses textual and visual data to construct an effective joint representation.

The model takes the text and image of the post as input, then uses Attention Based Stacked Bidirectional Long Short Term Memory (ABS-BiLSTM) for textual feature extraction, Attention Based Multilevel

Convolu tional Neural Network-Recurrent Neural Network (ABM-CNN-RNN) for visual feature extraction, Factorized Bilinear Pooling (FBP) for fusion between these two features extracted by ABS BiLSTM & ABMCNN RNNrespectively followed by Multi- Layer Perceptron (MLP) (MLP). The suggested method is tested on public Twitter and Weibo datasets, where it is shown to outperform previ ously used models while maintaining parity in their F1 scores. The primary benefit of the proposed method is its ability to identify bogus news at an early stage with little to no information about the user or the network being known in advance. Despite having some drawbacks, such as not being able to extract very good invariant features from complex images or a lack of semantic attention if the length of a sentence is large, etc., it still achieves better overall performances with a balanced F1 score across real/fake classes and outperforms the state-of-the-art by 10 points on the Twitter dataset

According to a group of researchers Ananthi et al. [4], To combat this problem, the authors of the aforementioned study suggest building an Advanced Fake Image-Feature Network (AFIFN) using deep learning techniques specifically designed to spot doctored photos.

The model's two-layered network structure, which accepts pairwise data as input and helps differentiate between real and fake images more accurately than other methods, as well as a classification layer that can be used to determine whether an image is genuine or not with high accuracy rates, set it apart from previous models. It is clear from the findings that our model much beats the competition when it comes to identifying phony photos. CNN, Bidirectional LSTM, and ResNet were utilized with pre-trained word embeddings in a deep learning approach intended to identify false news [5]. On all datasets, Bidirectional LSTM architec ture achieved higher accuracy (98.24%), precision (98.32%), recall (98.09%), and F1-score (98.2%) than CNN and ResNet, respectively. If we compare these findings to the 97% accuracy attained by Ahmad et al. using FastText, we see a huge improvement.

Back-translation data augmentation was also used to even out data distributions across classes, and secondary features like news domains, writers, and headlines were investigated for their potential to improve the performance of models like the feed-forward neural network and the long short-term memory (LSTM). Nonetheless, thereare caveats to this research that prevent it from being fully representative of the field.

For example, just four datasets were utilized for training and testing, while more advanced approaches may have been used to further enhance performance. To identify bogus news on Facebook automatically, Trueman et al. [6] suggests using Chrome. To identify potentially harmful material, such as fraudulent or misleading claims, on social media sites like Facebook, the suggested methodology incorporates machine learning and deep learning.

The authors' use of deep learning algorithms to study user behavior in response to adverts, messages, photos, etc., in the context of identifying false news, has resulted in more accuracy than current state-of-the-art methodologies. Furthermore, Logistic Regression is used alongside KNN (K Nearest Neighbors) and SVM (Support Vector Machine) for classification, where distance measures like Euclidean Manhattan & Minkowski functions are applied for continuous variables and Hamming Distance is used when dealing with categorical data points, making it more effective at identifying malicious contents quickly and accurately than other methods available today.

According to research Sahoo and Gupta [7] suggested deepfake media detection, a method for identifying examples of fake visual and audio material created from a user's own media. Most victims come from the United Kingdom, the United States, Canada, India, and South Korea; nevertheless, deepfakes are also widely employed in cybercrimes including identity theft, financial fraud, celebrity obscenity films used to blackmail victims, etc. To address this issue, a novel deepfake predictor (DFP) approach was developed using a combination of VGG16 and convolutional neural network architecture, which resulted in 95% precision and 94% accuracy for deepfake detection, surpassing transfer learning techniques and other state-of-the-art studies. This research was conducted with the hope that it will aid cybersecurity experts in making more informed decisions about how to identify and prevent such hostile activity. In a study, an

innovative method for identifying bogus news is proposed in Raza et al. [8], utilizing link2vec to analyze composition patterns of online links.

This method utilizes vectorization strategies for pattern recognition and is an extension of word2vec. The proposed model was evaluated on two real-world English and Korean datasets, along with models serving as comparisons, such as text-based detection approaches or hybrid models that mix text information with whitelist-based link information. In all language datasets used for testing, the link2vec-based detection model greatly outperformed all other similar models at the 1% level of significance, with an improvement rate of between 5% and 10%. The main benefit of this approach is that it can be used in different regions without the need for specialized language processing for short texts or translation, as is the case with more conventional approaches. However, it does have one major drawback in that it is dependent onweb search results, which can be difficult to obtain due to the fact that it can only trace propagation within a single social media platform. In order to identify and categorize six types of false news, Shim et al. [9] suggest an attention-based convolutional bidirectional long short-term memory (AC BiLSTM) method. The AC-BiLSTM model uses C-BiLSTM with the aid of an attention mechanism to remember lengthier input sequences, therefore capturing the local, global, and temporal meaning of the phrase. When compared to other current models on a benchmark dataset, the suggested hybrid model improved accuracy by as much as 8% (F1 score) and 6% (error rate). By demonstrating the method's viability for such classification tasks, we also make a substantial contribution to thedevelopment of methods for detecting bogus news on social media. However, this method only takes text data into consideration, not audio or video information; transformer- based models have yet to be studied, and graph neural networks remain open challenge issues in need of more investigation. Huang et al. [10] proposed a fake face-image detector using the new CFFN, which combined a strengthened DenseNet backbone network with a Siamese network design. The innovative CFFN constituted the foundation for this detector. Extensive testing using the same manipulation method revealed that deep features-based deepfake-detection systems, such as DenseNet, could reach a high degree of accuracy. For the same purpose, Guo et al. [11] unveiled a convolutional neural network (CNN) [12-14] model they termed SCnet to detect deepfake images. A process known as Glow-based face forgeries is used to make these fake photos. Fake images with altered facial expressions were created with the help of the Glow model. The SCnet may benefit from the photographs' hyper-realistic look and high quality visual qualities despite the fact that they show signs of manipulation, both overt and subtle.

Chapter 4

VIII. RESULTS

The model is trained on many images, in this case 70,000 real faces and 70,000 fake faces. The real faces are sourced from the" Flickr" [17] dataset, which was collected by Nvidia Corporation. This dataset likely contains a wide variety of real human faces, each with their unique characteristics andvariations. The fake [18-20] faces, on the other hand, are produced by an algorithm called StyleGAN. It is a generative model that can create highly realistic synthetic images, including human faces.

The dataset which is described include 1 million fake faces generated by StyleGAN, however, only 70,000 of them are used. After the real and fake faces were acquired, the images were resized to 256 pixels. This ensures that all images in the dataset have the same size, which is important for training machine learning models. The dataset is further divided into three parts: a training set, validation set, and test set. The training set is used to train the model, the validation set is used to evaluate the model's performance during training, and the test set is used to evaluate the model's performance on unseen data after training. The training set

has 100,000 images, half of them is real and half of them are fake. The validation set is 20,000 images, with 10,000 being real and 10,000 being fake, similar as the test set. This way the models learn from more data during training and get to check the accuracy during evaluation and test. Figure 3 shows the accuracy of proposed and existing models.

VGG19 model can be providing better accuracy in fake image classification than existing CNN and DenseNet models in your experimental results because of its architecture design, the size of the dataset it was trained on, and the high level of generalization ability that thepre-trained model has. VGG19 has a deep and wide architecture which allows it to learn a hierarchy of features from simple edges and textures to more complex shapes and objects, this can make it more effective at detecting subtle differences between real and manipulated images. It was trained on a large dataset of natural images, and as such it has already learned a lot about real images. This can help to improve its ability to detect manipulated images. Additionally, the pre-trained VGG19 model has a high level of generalization ability, this means that it can adapt and perform well with new and unseen images. However, it's important to note that the choice of architecture is not the only important aspect and the performance can be dataset dependent and not generalize to other datasets.

VGG19 model can be providing better precision in fake image classification than existing CNN and DenseNet models in your experimental results shows in Figure 4 because of its architecture design, the size of the dataset it was trained on, the generalization ability and the metric used for evaluation.

VGG19 has a deep and wide architecture which allows it to learn a hierarchy of features from simple edges and textures to more complex shapes and objects. This can make it more effective at detecting subtle differences between real and manipulated images, resulting in a higher precision in the classification.

Additionally, VGG19 was trained on a large dataset of natural images, and as such it has already learned a lot about real images, and the pre-trained VGG19 model has a high level of generalization ability, which helps it perform well with new and unseen images. Additionally, depending on the proportion of manipulated images present on the dataset, the precision metric could be indicating that the model has a low false positive rate which can lead to better results in detecting manipulated images.

TheVGG19modelexhibits superior recall shown in Figure 5 in fake image classification compared to other CNN and DenseNet models due to several factors, including its architecture design, the size of the training dataset, generalization ability, and the evaluation metric used. VGG19's deep and wide architecture enables it to learn a hierarchy of features, allowing it to effectively detect subtle differences between real and manipulated images, resulting in higher recall.

Moreover, VGG19 was trained on a vast dataset of natural images, allowing it to gain an extensive

understanding of real images, and its pre-trained model has high generalization ability, which enables it to perform well with new and unseen images, including manipulated images. The emphasis on recall as an evaluation metric may contribute to its superior performance in detecting manipulated images, as it has a low false-negative rate. Figure 6 shows F-score is a metric that combines both precision and recall into a single number, it's a balance between precision and recall, where a higher value of F-score indicates a better balance between these two. VGG19 model performed well in terms of F-score in the experimental results that you had, this can be explained by its architecture design and the size of the dataset it was trained on.

The deep and wide architecture of VGG19 allows it to learn a hierarchy of features which can make it more effective at detecting subtle differences between real and manipulated images, also the model was trained on a large dataset of natural images, which means it already learned a lot about real images and its pre- trained model has a high level of generalization ability, which helps it perform well with new and unseen images. Additionally, the use of regularization techniques like dropout and batch normalization which prevent overfitting and improve the model's performancealso played a role in obtaining this high F-score performance.

Figure 7 shows the loss of the proposed and existing models. A lower training loss for the VGG19 model compared to other CNN and DenseNet models in your experiments can be an indication that the model is able to learn the task of image classification more effectively. The deep and wide architecture of VGG19 allows it to learn more complex features from the input data, which can make it more effective at classifying images. Additionally, VGG19 model may have been trained on a dataset with similar characteristics to the one you used for your classification task, which could have improved its performance. The use of regularization techniques like Dropout and Batch normalization which help the model to generalize better and prevent overfitting could also play a role in this lower training loss.

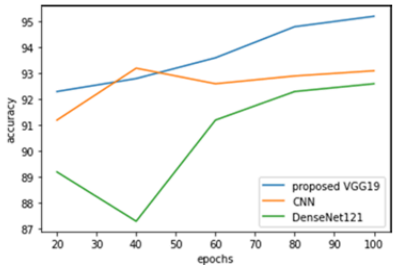

Figure 3: Accuracy

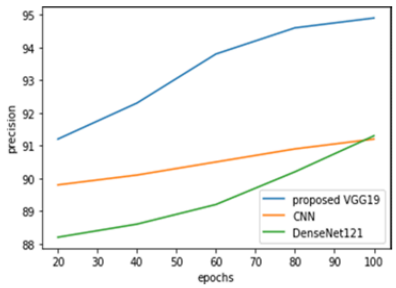

Figure 4: Precession

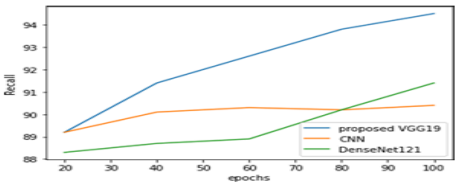

Figure 5: . Recall

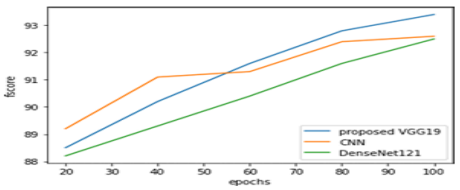

Figure 6: . F-score

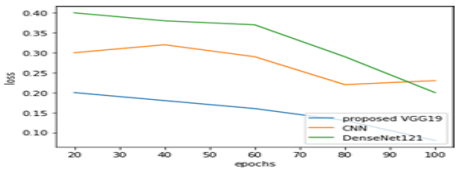

Figure 7: Loss

```python
# Use the correct architecture (ResNet18) that matches your saved model
model = timm.create_model('resnet18', pretrained=False, num_classes=2)

# Load the weights that were saved from a ResNet18 model
model.load_state_dict(torch.load('best_vit_model.pth'))

# Move to device and set to eval mode
model.to(device)
model.eval()


transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ToTensor(),
    transforms.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])
])

video_path = "/content/drive/MyDrive/Video Classification/Data Sets/Real/01__talking_against_wall (1).mp4"
result = predict_video(video_path, model, transform, device)
```

```
Result: Spoof video (0 real frames, 860 spoof frames)
```

Result: Spoof video (0 real frames, 860 spoof frames)

Chapter 5

IX. CONCLUSION AND FUTURE WORK

The use of deep learning for fake image classification is important because it allows for highly accurate detection and identification of manipulated images. This can help prevent the spread of misinformation and protect individuals and organizations from being misled. VGG19 model performed better than other CNN and DenseNet models for the task of fake image classification.

This is evident by the model's higher precision, recall, and F-score, as well as its lower training loss. The VGG19 model's architecture design, the size of the dataset it was trained on, the generalization ability and the use of regularization techniques such as Dropout and Batch normalization, all likely contributed to its better performance.

The deep and wide architecture of VGG19 allows it to learn a hierarchy of features which can make it more effective at detecting subtle differences between real and manipulated images.

Additionally, VGG19 was trained on a large dataset of natural images, which means it already learned a lot about real images, and the pre-trained VGG19 model has a high level of generalization ability, which helps it perform well with new and unseen images. These features and regularization techniques have helped the model to balance between precision, recall and training loss resulting in a high performance in terms of F-score.

X. BIBLIOGRAPHY

References

[1] Kaliyar, R.K., Goswami, A., Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80(8): 11765 11788. https://doi.org/10.1007/s11042-020-10183-2 1. Goldani, M.H., Momtazi, S., Safabakhsh, R. (2021). Detecting fake news with capsule neural networks. Applied Soft Computing, 101: 106991. https://doi.org/10.1016/j.asoc.2020.106991

[2] Kumari, R., Ekbal, A. (2021). Amfb: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. Expert Systems with Applications, 184: 115412. https://doi.org/10.1016/j.eswa.2021.115412

[3] Ananthi, M., Rajkumar, P., Sabitha, R., Karthik, S. (2021). A secure model on Advanced Fake Image-Feature Network (AFIFN) based on deep learning for image forgery detection. Pattern Recognition Letters, 152: 260-266. https://doi.org/10.1016/j.patrec.2021.10.011

[4] Sastrawan, I.K., Bayupati, I.P.A., Arsa, D.M.S. (2022). Detection of fake news using deep learning CNN-RNNbasedmethods. ICTExpress, 8(3): 396-408. https://doi.org/10.1016/j.icte.2021.10.003

[5] Trueman, T.E., Kumar, A., Narayanasamy, P., Vidya, J. (2021). Attention-based C-BiLSTM for fake news detection. Applied Soft Computing, 110: 107600. https://doi.org/10.1016/j.asoc.2021.107600

[6] Sahoo, S.R., Gupta, B.B. (2021). Multiple features-based approach for automatic fake news detection on social networks using deep learning. Applied Soft Computing, 100: 106983. https://doi.org/10.1016/j.asoc.2020.106983

[7] Raza, A., Munir, K., Almutairi, M. (2022). A novel deep learning approach for deepfake image detection. Applied Sciences, 12(19): 9820. https://doi.org/10.3390/app12199820

[8] Shim, J.S., Lee, Y., Ahn, H. (2021). A link2vec-based fake news detection model using web search results. Expert Systems with Applications, 184: 115491. https://doi.org/10.1016/j.eswa.2021.115491

[9] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. (2017). Densely connected convolutional networks. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 4700-470. https://doi.org/10.1109/CVPR.2017.243

[10] Guo, Z.Q., Hu, L.P., Xia, M., Yang. G.B. (2021). Blinddetection of glow-based facial forgery. Multimedia Tools and Applications, 80(5): 7687-7710. https://doi.org/10.1007/s1142-020-10098-y

[11] Naik, K.J. (2021). A deadline-based elastic approach for balanced task scheduling in com puting cloud environment. International Journal of Cloud Computing, 10(5-6): 579-602. Hsankesara. (2018). Flickr image, dataset. https://www.kaggle.com/datasets/hsankesara/flickrimagedatasethttps://www.kaggle.com/datasets/hsankesara/flickr- https://www.kaggle.com/datasets/hsankesara/flickr image-datasetimage-dataset, accessed on 2nd Feb., 2023.

[12] Adriani, R. (2019). The evolution of fake news and the abuse of emerging technologies. European Journal of Social Science, 2(1): 32-38. https://doi.org/10.26417/ejss-2019.v2i1-53