Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220655

Deep fake detection using deep learning techniques

Gayathri S[1,3], Santhiya S[2], Nowneesh T[1], Sanjana Shuruthy K[1] and Sakthi S[1]

[1]Computer Science and Engineering, Kongu Engineering College, Erode, India
[2]Artificial Intelligence, Kongu Engineering College, Erode, India

[3]sgayathricse97@ gmail.com

Abstract. Deep fake is the artificial manipulation and creation of data, primarily through photographs or videos into the likeness of another person. This technology has a variety of applications. Despite its uses, it can also influence society in a controversial way like defaming a person, Political distress, etc. Many models had been proposed by different researchers which give an average accuracy of 90%. To improve the detection efficiency, this proposed paper uses 3 different deep learning techniques: Inception ResNetV2, EfficientNet, and VGG16. These proposed models are trained by the combination of Facforensic++ and DeepFake Detection Challenge Dataset. This proposed system gives the highest accuracy of 97%.

Keywords: deepfake detection, inception resNetv2, EfficientNet B4, VGG16, FaceForensic++, DeepFake Detection Challenge Dataset.

1. Introduction

Deepfake is an artificial intelligence technique for creating convincing pictures, audio, and videos forgeries. These deepfake are generated using a deep learning model known as Generative Adversarial Networks (GAN). There are two neural networks in this system: Generators and Detectors. As the name implies, generators are responsible for creating new images from existing ones, while detectors are responsible for determining whether the images are real. Both networks are continually interacting with one another to create a better network. It has vast applications in science and arts. Although it has a distressing impact on society. A health charity in the United Kingdom, for example, utilized a deepfake to have David Beckham give an anti-malaria message. This communication was also sent out in nine different languages. The most famous and deadly application of technology is when others choose to utilize it for bad purposes. They could be used to disseminate false information from a reputable source, such as election propaganda. To avoid this, a detection methodology is needed. Watermarking, media verification markers, signatures, and chain-of-custody recording are examples of methods that help prove integrity across the media lifecycle. Because it checks and tracks integrity throughout the content lifecycle or verifies it at the distribution endpoint, authentication is the most effective approach to avoid misleading modification of trustworthy media. Media provenance refers to solutions that disclose information about the media's origin, either directly in the media or as metadata. A reverse media search can also be a useful provenance tool; a list of websites where a given media has previously appeared can be used to prove the media's origin. Provenance, when combined with authenticity, can be useful forensic tools in the fight against deep fakes. Deep fake detection solutions use multimodal detection approaches to identify whether target media has been modified or generated

DOI: 10.54254/2755-2721/2/20220655

synthetically. Existing detection methods are classified into two types: manual and algorithmic. Human media forensic practitioners, typically armed with digital tools, use manual procedures. Algorithmic detection identifies corrupted media using an AI-based algorithm. An algorithmic detection approach is employed to to discover the deepfake because a human method may result in incorrect predictions. Three different models have been identified for detecting deep fake media: Inception ResNet V2, EfficientNet, and VGG16. The combinations of both Faceforensic++ and Deepfake detection Challenge Dataset have been used for identifying deep fake in videos.

2. Literature review

These videos have both benefits and harm to society. This deep fake can be incorporated into surveillance such as facial reconstruction for high authority officials to enter the private sectors safely and securely. This technology is incorporated in many tech giant companies like Face book, Apple, etc.[1]. On other hand, it may also lead to political distressing, defaming a celebrity in public. Some technologies are involved in modifying lips and eye movement to simulate the presence of another person who speaks coherently to the lip movement [2]. Some imply optical flow-based techniques for extracting the video that adopts transfer learning techniques on a part of the dataset while others have been used fine-tuning. Then they are subjected to deep fake detection using CNN [3]. In most of the research, CNN has been used for feature extraction regardless of the format of the dataset (either video, audio, or images). Few types of research implied CNN as a feature extraction [4, 5]. Deep fake is an artificial intelligence technique for creating convincing pictures, audio, and videos forgeries. These deep fake are generated using a deep learning model known as Generative Adversarial Networks (GAN). There are two neural networks in this system: Generators and Detectors [6].One of the researches includes CNN and RNN for detecting deep fake images. The characteristics in the video were extracted using CNN [7, 8]. These extracted data are then used for training the Recurrent Neural Network (RNN) to determine whether the video is subjected to manipulation [9]. Artifact extraction on images that were made during deep fake creation was used to train the model. This has a great impact on the model effect as it needs low resources and it is a highly robust model [10]. Deep fake was a term coined from "deep learning" and "fake" inferring that media uses the deep learning method for generating deep fake images and videos [11, 12]. For this, much research has been undergone. Researchers surveyed the methods to be used for deep fake detection [13][14]. These models include classical machine learning methods, deep learning methods, block chain-based techniques, and statistical techniques [15][16]. Among these, the Deep learning model shows good results over others. Spatiotemporal methods have also been used for deep fake detection [17][18]. A sequence of frames from the videos was extracted using sliding window methods. With the help of these images, this system analyses the intraframe spatial dependencies and inter frame temporal dependencies which makes better recognization of the deep fakes overtaking a single frame and analyzing it using machine learning models [19][20].

3. Dataset description

For every Deep learning method, a large collection of datasets is needed to run a model which makes it learn efficiently. For training the model, proposed system choose two different types of datasets: FaceForensics++ and Deep Fake Detection Challenge dataset. FaceForensic++ is an open-source dataset that has a wide collection of different facial reconstruction videos like Face2Face, Deepfake, Face Swap, and Neural Textures along with the original sequence. Each category has 1000 videos. Another dataset called the Deepfake Detection Challenge dataset has train and test videos along with two files, metadata.json, and submission.csv. Each train and test dataset has 400 videos. The metadata.json has information regarding the videos as to whether the videos are real or fake. Submission.csv file analyzes or evaluates the model's performance, which has the originality of the video (either fake or real) represented in 0's and 1's (0 - fake and 1- real). The reason for using two datasets for detecting fake videos is to improve the level of identifying the fake images among the original ones. The model

DOI: 10.54254/2755-2721/2/20220655

has been learned from different faces and classifies itself with the videos. Table 1 shows the number of files in both the dataset.

Table 1. Original and deep fake dataset.

DataSet	Video Count
FaceForensics++	
Original	1000
DeepFake	1000
Deepfake Detection Challenge Dataset	
test_videos	400
Train_sample_videos+Metadata.json	400 + 1
Submission.csv	1

Deepfake videos and its original videos have been taken from Faceforensics++ and both test_videos and Train_sample_videos as a new dataset. This dataset has been divided into two parts: 80 percent for training and 20 percent for testing and validation.

4. Proposed system

The Proposed work mainly focuses on the originality of videos whether it is deepfake or real. Since the selected dataset consists of videos, it has been difficult for the model to train as a raw input. Proposed system implemented by extracting the facial features from the videos that have been given to the model for training. get_frontal_face_detector() from dlib have been used to identify the face in the video. Additionally, OpenCV module to extract the image according to the height and width obtained from get_frontal_face_detector().

Deep Learning is a type of machine learning that can perform cognitive processes similar to those performed by humans, such as learning by example [21][22]. It has been used in a variety of upcoming technologies, including self-driving automobiles, voice recognition in phones and tablets, and so on [23]. Text, image, and voice data were used to train a classification model using deep learning [24]. Deep Learning does classification using a vast amount of labeled data [25]. Neural networks, which include as many as 150 hidden layers, are commonly used in deep learning approaches [26]. Deep learning achieves improved accuracy, sometimes surpassing human ability, with the use of tagged data and neural networks [27]. Inception ResNet V2, EfficientNet, and VGG16 models have been used for identifying DeepFake images [28].

DOI: 10.54254/2755-2721/2/20220655

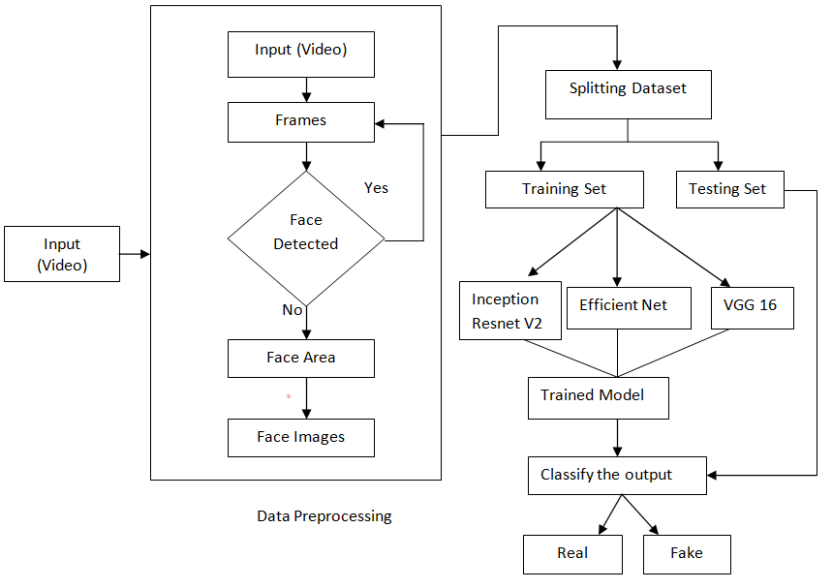

Figure 1. Architecture diagram.

4.1. Inception, resnet, V2

The convolutional neural network Inception-ResNetv2 was trained on over 1,000,000 photos from the ImageNet Dataset. The 164-layer network is capable of distinguishing between 1000 different object kinds, including mice, keyboards, and other animals. As a consequence, the Network has amassed a range of rich representations of characteristics for a variety of photos 299-by-299 pictures are put into the network, which produces a table of predicted class probabilities. It is built on the Residual connection and the Inception structure. The Inception-Resnet block combines convolutional filters of various sizes with residual connections. The incorporation of residual connections not only enhances the end product's quality but also tackles the issue of deep structural deterioration and lowers training time in half. Residual and Inception connections are combined in the Inception-ResNet architecture. Fig.2 explains the blocks in Inception ResNet V2 [29].

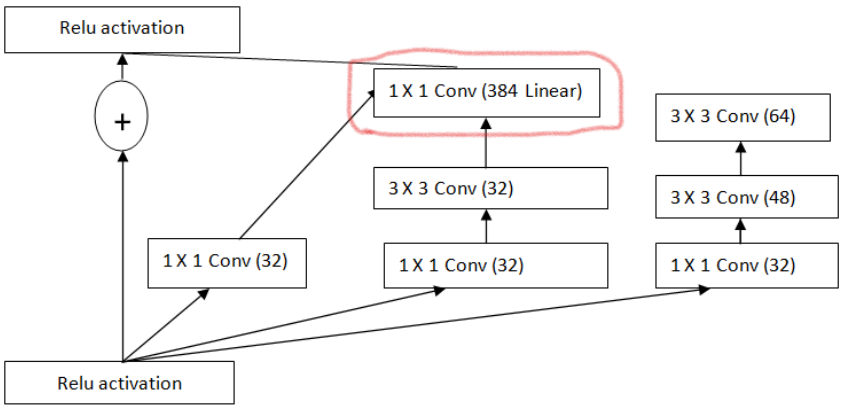

Figure 2. Blocks of residual inception.

DOI: 10.54254/2755-2721/2/20220655

1. After each Inception block, a filter expansion layer (1 convolution without activation) is added before increasing the filter bank's dimensionality then it is added to match the depth of the input.
2. Batch-normalization is only applied to the standard layers by Inception-ResNet, but not to the summations.

After a few tens of millions of repetitions, the leftover variations began to show instabilities, and the network simply "died" early in the training, with the last layer before the average pooling producing just zeros. This could not be avoided, regardless of whether the learning rate was reduced or an additional batch normalization layer was added. Scaling down residuals before adding them to the preceding layer activation appears to help stabilize the training, according to them. The residuals were scaled using the scaling of factors ranging from 0.1 to 0.3

Inception ResNet V2 and EfficientNet are trained using the classified images. The classified real and fake images have been given as input to the proposed models. The models train themselves with the classified images through pools of layers and give better accuracy over the existing model. It is trained as a supervised model, where the training includes the real and fake classification of images. With the help of those images, the proposed models are trained using binary classification types. The model is saved for later purposes as an h5 file. With that file, test the model against the test video which includes the extraction of images and predicting them as if it is a real or fake.

5. Results and discussion

The results of the implemented models are measured using accuracy and precision. During Implementation, InceptionResNetV2 gives an accuracy of 97 percent for the extracted images from the videos. While, EffeicientNet B4 was used to predict fake images with an accuracy of 96 percent. On the other hand, VGG16 gives an accuracy of 95 percent. Table 2 gives the overall accuracy of the model proposed.

Table 2. Accuracy of the model.

Model	InceptionResNetV2	EfficientNetB4	VGG 16
Accuracy	97	96	85

It is trained and measured the performance using 2 different optimizers: SGD and Adam.

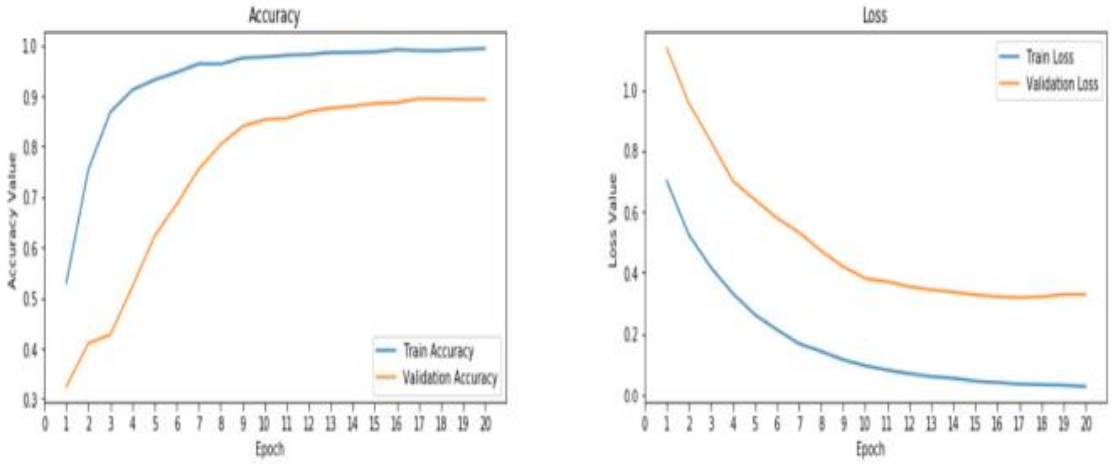

Figure 3. Inception ResNet V2 Train Accuracy (a) and Loss (b) for Adam optimizer.

DOI: 10.54254/2755-2721/2/20220655

In Fig. 3, For the Adam Optimizer, The model shows the maximum training accuracy of .99 and minimum loss of 0.1 whereas for validation it has the maximum accuracy of 0.9 and minimum loss of 0.38. Fig. 4 demonstrate the SGD optimizer used for training Inception ResNet V2, with a training accuracy of .99 and minimum loss of 0.01.

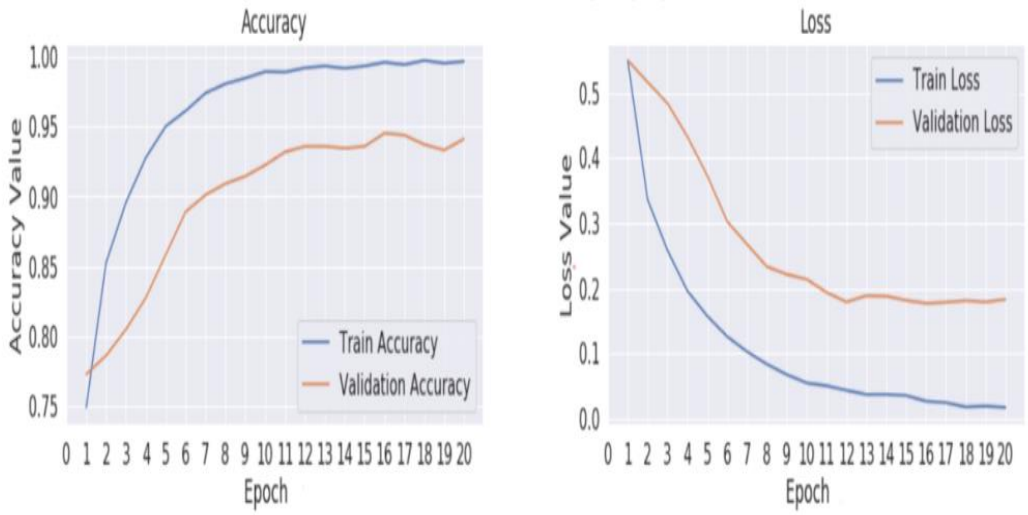

Figure 4. Accuracy (a) and validation (b) for inception resNet V2 during training using SGD optimizer.

Fig. 5 and Fig. 6 shows the difference between model training and loss under different optimizers for VGG 16. Both the Graph shows the maximum accuracy the model attains nearly .84 and minimum loss of .3 to .35

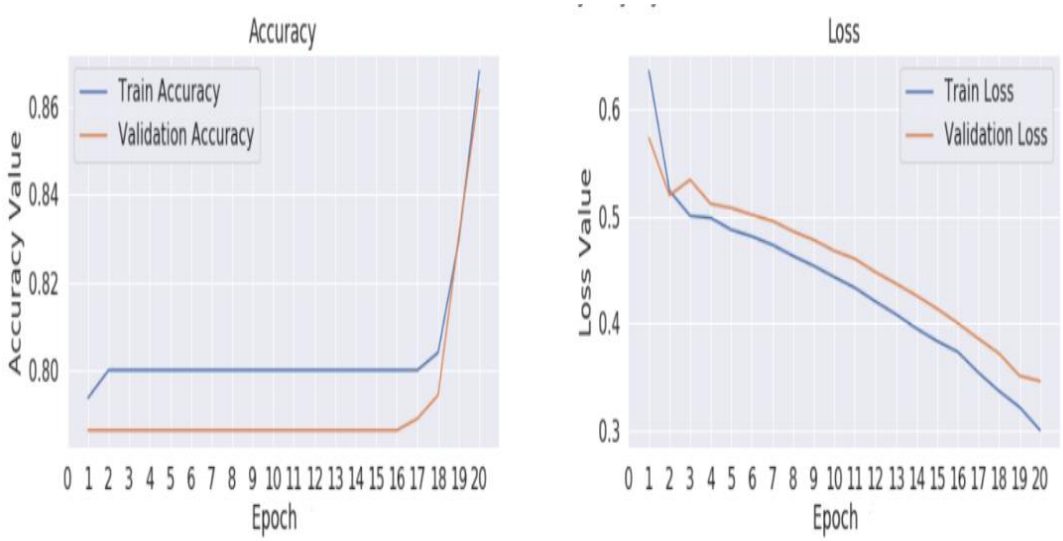

Figure 5. VGG16 Training (a) and Loss (b) graph for SGD optimizer.

DOI: 10.54254/2755-2721/2/20220655

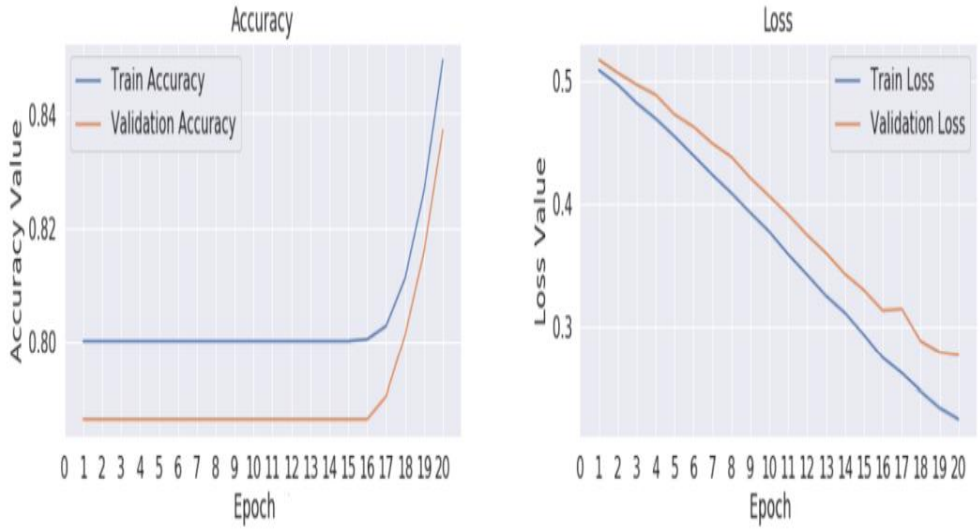

Figure 6. VGG16 Training (a) and Loss (b) graph for Adam Optimizer.

Accuracy is the total ratio of classified correct positive and negative example to the whole classified example. Table 3 shows the optimizer accuracy comparison between the models

Table 3. Accuracy of the model for the optimizers.

Optimizer	Adam	SDG	Ada Grad
Inceptionresnetv2	97.6	98.3	96.0
VGG 16	84.2	89.9	86.3
Efficient Net	94.8	95.2	95.3

Precision is the number of correctly classified positive examples to the total correctly classified examples. Table 4 shows the optimizer precision comparison between the models.

Table 4. Precision of the model for the optimizers.

Optimize r	Adam	SDG	AdaGrad
Inception resnetv2	89.4	86.3	92.7
VGG 16	72.8	74.8	79.6
Efficient Net	86.4	87.1	84.3

Recall is the number of correctly classified positive examples to the total positive prediction. Table 5 shows the optimizer recall between the models. Table 6 shows the detailed training accuracy of each model at different intervals of epochs.

Table 5. Recall of the model for the optimizers.

Optimizer	Adam	SDG	AdaGrad
InceptionResnetv2	87	90	91.1
VGG 16	73.8	75.5	79.8
EfficientNet	89.4	92.6	90.7

DOI: 10.54254/2755-2721/2/20220655

Table 6. Detailed training accuracy of each model at different intervals of epochs.

Epochs	InceptionResnetV2	EfficientNet	VGG16
5	95.1	92.3	80
10	98.1	95.6	80.1
15	99.2	98.4	86.4
20	99.5	99.3	93.2
25	99.7	99.9	94.8

6. Conclusion

DeepFake evolving around the world in different ways. Detecting fake images would be a difficult task in the upcoming years. Misusing technology leads to misery among people. Deep Learning is the technology that involves in generation and detection of fake images. Fake images have been extracted from captured videos using dlib library. The proposed work implemented three different deep learning algorithms such as VGG16, EfficientNet, and Inception ResNet V2. InceptionResnetV2 results higher accuracy with minimum error rate. Implementation can be further extended using preprocessing and different deep learning models in future. It would be helpful in the identification of real and fake images in a better way.

References

[1] Ahmed, S. R. A., & Sonuç, E., ‘‘Deepfake detection using rationale-augmented convolutional neural network”, Applied Nanoscience, pp. 1-9. (2021).

[2] Amerini, Irene, et al. ‘‘Deepfake video detection through optical flow based cnn. ’’ Proceedings of the IEEE/CVF international conference on computer vision workshops. (2019).

[3] Chang, Xu, Jian Wu, Tongfeng Yang, and Guorui Feng. ‘‘Deepfake face image detection based on improved VGG convolutional neural network. ’’In 2020 39th chinese control conference (CCC), IEEE, pp. 7252-7256. (2020).

[4] Hsu, Chih-Chung, Yi-Xiu Zhuang, and Chia-Yen Lee. ‘‘Deep fake image detection based on pairwise learning. ’’Applied Sciences vol. 10, no. 1, pp. 370. (2020).

[5] Guarnera, Luca, Oliver Giudice, and Sebastiano Battiato. ‘‘Deepfake detection by analyzing convolutional traces. ’’ In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 666-667. (2020).

[6] Güera, David, and Edward J. Delp. ‘‘Deepfake video detection using recurrent neural networks. ’’ In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, pp. 1-6. (2018).

[7] Ismail, Aya, Marwa Elpeltagy, Mervat S. Zaki, and Kamal Eldahshan. ‘‘ A New Deep Learning-Based Methodology for Video Deepfake Detection Using XGBoost.’’ Sensors vol. 21, no. 16, pp. 5413. (2021).

[8] Jung, Tackhyun, Sangwon Kim, and Keecheon Kim. ‘‘Deepvision: Deepfakes detection using human eye blinking pattern.’’ IEEE Access 8, pp. 83144-83154. (2020).

[9] Li, Yuezun, and Siwei Lyu. ‘‘ Exposing deepfake videos by detecting face warping artifacts. ’’ arXiv preprint arXiv:1811.00656, (2018).

[10] Malolan, Badhrinarayan, Ankit Parekh, and Faruk Kazi. ‘‘Explainable deep-fake detection using visual interpretability methods. ’’ In 2020 3rd International Conference on Information and Computer Technologies (ICICT), IEEE, pp. 289-293. (2020).

[11] Mitra, Alakananda, Saraju P. Mohanty, Peter Corcoran, and Elias Kougianos. ‘‘A machine learning based approach for deepfake detection in social media through key video frame extraction. ’’SN Computer Science 2, no. 2, pp. 1-18. (2021).

[12] Sathishkumar V E, Changsun Shin, Youngyun Cho, “Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city”, Building Research & Information, Vol. 49. no. 1, pp. 127-143, 2021.

DOI: 10.54254/2755-2721/2/20220655

[13] Sathishkumar V E, Youngyun Cho, “A rule-based model for Seoul Bike sharing demand prediction using Weather data”, European Journal of Remote Sensing, Vol. 52, no. 1, pp. 166-183, 2020.
[14] Sathishkumar V E, Jangwoo Park, Youngyun Cho, “Seoul Bike Trip duration prediction using data mining techniques”, IET Intelligent Transport Systems, Vol. 14, no. 11, pp. 1465-1474, 2020.
[15] Sathishkumar V E, Jangwoo Park, Youngyun Cho, “Using data mining techniques for bike sharing demand prediction in Metropolitan city”, Computer Communications, Vol. 153, pp. 353-366, 2020.
[16] Sathishkumar V E, Yongyun Cho, “Season wise bike sharing demand analysis using random forest algorithm”, Computational Intelligence, pp. 1-26, 2020.
[17] Sathishkumar, V. E., Wesam Atef Hatamleh, Abeer Ali Alnuaim, Mohamed Abdelhady, B. Venkatesh, and S. Santhoshkumar. "Secure Dynamic Group Data Sharing in Semi-trusted Third Party Cloud Environment." Arabian Journal for Science and Engineering (2021): 1-9.
[18] Chen, J., Shi, W., Wang, X., Pandian, S., & Sathishkumar, V. E. (2021). Workforce optimisation for improving customer experience in urban transportation using heuristic mathematical model. International Journal of Shipping and Transport Logistics, 13(5), 538-553.
[19] Pavithra, E., Janakiramaiah, B., Narasimha Prasad, L. V., Deepa, D., Jayapandian, N., & Sathishkumar, V. E., Visiting Indian Hospitals Before, During and After Covid. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 30 (1), 111-123, 2022.
[20] Easwaramoorthy, S., Moorthy, U., Kumar, C. A., Bhushan, S. B., & Sadagopan, V. (2017, January). Content based image retrieval with enhanced privacy in cloud using apache spark. In International Conference on Data Science Analytics and Applications (pp. 114-128). Springer, Singapore.
[21] Sathishkumar, V. E., Agrawal, P., Park, J., & Cho, Y. (2020, April). Bike Sharing Demand Prediction Using Multiheaded Convolution Neural Networks. In Basic & Clinical Pharmacology & Toxicology (Vol. 126, pp. 264-265). 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY.
[22] Subramanian, M., Shanmuga Vadivel, K., Hatamleh, W. A., Alnuaim, A. A., Abdelhady, M., & VE, S. (2021). The role of contemporary digital tools and technologies in Covid-19 crisis: An exploratory analysis. Expert systems.
[23] Babu, J. C., Kumar, M. S., Jayagopal, P., Sathishkumar, V. E., Rajendran, S., Kumar, S., ... & Mahseena, A. M. (2022). IoT-Based Intelligent System for Internal Crack Detection in Building Blocks. Journal of Nanomaterials, 2022.
[24] Subramanian, M., Kumar, M. S., Sathishkumar, V. E., Prabhu, J., Karthick, A., Ganesh, S. S., & Meem, M. A. (2022). Diagnosis of retinal diseases based on Bayesian optimization deep learning network using optical coherence tomography images. Computational Intelligence and Neuroscience, 2022.
[25] Liu, Y., Sathishkumar, V. E., & Manickam, A. (2022). Augmented reality technology based on school physical education training. Computers and Electrical Engineering, 99, 107807.
[26] Sathishkumar, V. E., Rahman, A. B. M., Park, J., Shin, C., & Cho, Y. (2020, April). Using machine learning algorithms for fruit disease classification. In Basic & clinical pharmacology & toxicology (Vol. 126, pp. 253-253). 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY.
[27] Sathishkumar, V. E., Venkatesan, S., Park, J., Shin, C., Kim, Y., & Cho, Y. (2020, April). Nutrient water supply prediction for fruit production in greenhouse environment using artificial neural networks. In Basic & Clinical Pharmacology & Toxicology (Vol. 126, pp. 257-258). 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY.
[28] Sathishkumar, V. E., & Cho, Y. (2019, December). Cardiovascular disease analysis and risk assessment using correlation based intelligent system. In Basic & clinical pharmacology &

DOI: 10.54254/2755-2721/2/20220655

toxicology (Vol. 125, pp. 61-61). 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY.

[29] Kotha, S. K., Rani, M. S., Subedi, B., Chunduru, A., Karrothu, A., Neupane, B., & Sathishkumar, V. E. (2021). A comprehensive review on secure data sharing in cloud environment. Wireless Personal Communications, 1-28.