IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

DEEPFAKE DETECTION

[1]Rayapuraju Sreenitha, [2]Peri Abhinaya Deepika, [3]Rachakonda Meghna, [4]Gajula Venkat Sai, [5]A. Amara Jyothi

[1,2,3,4]Student, Department of CSE(AI&ML), CMR Engineering College, Hyderabad, Telangana

[5] Assistant Professor Department of CSE(AI&ML), CMR Engineering College, Hyderabad, Telangana

Abstract: In the evolution of artificial intelligence, especially in image manipulation techniques leads to creation of deepfakes, there are necessary needs for detection methods to strive against misinformation and maintain the honor of digital media. Convolutional Neural Networks (CNNs) became the most dominant tool in this regard, enables the identification of deepfake content by inspecting the patterns, textures, and peculiarity within the images.

Our project centralizes on suggesting the vigorous approach to deepfake detection, imposing neural networks to examine and compare manipulated elements against a inclusive dataset to authentic images. By precisely examining the differences indicative of manipulation, our algorithm can precisely notice the difference between original and deepfake media, thereby diminishing the misleading content.

In an era of advanced artificial intelligence manipulation, the development of an effective deep fake detection system is essential to protect the trust and integrity of digital media. Using algorithms and real-time monitoring, we're reducing the spread of fake artificial intelligence content. These algorithms are looking for inconsistencies and patterns which indicate that this may not be true. We use these advanced tools to keep an eye on misleading videos and images, so that we can stay alert and aware of what's out there online. It's as if we have a Digital Detective that helps us figure out what is real and what isn't, in order to be able to believe the things we see on the Internet.

IndexTerms **- Deepfake Detection, Authenticity Verification.**

I. INTRODUCTION

Deepfake detection is a crucial process aimed at noticing the difference between authentic and manipulated media, produced using advanced AI and deep learning techniques. It's main objective is to protect the nobility of digital information by tackling the spread of AI generated content and suppress the spread of misinformation. As the advancement of the technology continues, the capability to accurately identify fake images becomes increasingly important in ensuring that individuals can trust the information they come across online. By im- posing advanced algorithms and detection methods, deepfake detection plays a significant role in validating the reliability of social media platforms and maintaining public trust in the originality of online content. This dynamic approach is essential in diminishing the presumably harmful impacts of misinformation, thereby promoting a more informed and discerning digital society.

II. EXISTING SYSTEMS

2.1 Detecting false news

As per the study conducted by Zheng et al.(2018), the distinguishing of fake news and images is very tough. It has been put forward to study the problem of detecting false news, through an intensive inspection of forged news ,most constructive properties are determined from text words and pictures used in forged news. There are few unseen characteristics in words and images used in fake news, which can be perceived through a cluster of unseen properties acquired from this model through various layers. TI-CNN, named pattern has been put forward. By exhibiting the embedded characteristics in a unified space, TI-CNN is trained with a combination of text and image information at the same time.

2.2 Predict Fake Accounts

Raturi's 2018 architecture was proposed to identify counterfeit accounts in social networks, especially on Face- book. In this research, a machine learning feature was used to better predict fake accounts, based on their posts and the placement on their social networking walls. Support Vector Machine (SVM) and Complement Naıve Bayes (CNB) were used in this process, to validate content based on text classification and data analysis. The analysis of the data focused on the collection of offensive words, and the number of times they were repeated. For Facebook, SVM shows a 97 resolution where CNB shows 95 accuracy in recognizing Bag of Words (BOW) -based counterfeit accounts. The results of the study confirmed that the main problem related to the safety of social networks is that data is not properly validated before publishing.

2.3 Detect and Localize Fake Images

In 2017 study by Bunk et al , two systems were proposed to detect and localize fake images using a mix of resampling properties and deep learning. In the primary system, the Radon conversion of retesting of properties is set on superposed picture corrections. Deep learning classifiers and a Gaussian conditional domain pattern are then used to construct a heat map. In the next system, for identification and localization, software resampling properties are passed on overlapping object patches over a long-term memory (LSTM)- based network. To add-on, the detection/ localization production of both systems was inspected. The outcome proved that both systems are vital in detecting and settling digital image fraud.

2.4 Detect Manipulation and Fake images

According to Kim's and Lee's, digital forensics techniques are needed to detect manipulation and fake images used for illegal purposes. Thus, the researchers in this study have been working on an algorithm to detect fake im- ages through deep learning technology, which has achieved remarkable results in modern research. First, a converted neural network is applied to image processing. In addition, a high pass filter is used to get at hidden features in the image instead of semantic information in the image. For experiments, modified images are created using intermediate filter, Gaussian blurring, and added white Gaussian noise.

2.5 Detect Counterfeit News

Aphiwongsophon and Chongstitvatana, aimed to use automated learning techniques to detect counterfeit news. Three common techniques were used in the experiments: Naıve Bayes, Neural Network and Support Vector Machine (SVM). The normalization method is a major step to disinfect data before using the automatic learning method to sort information. The results show Naıve Bayes to have a 96.08 percent accuracy in detecting counterfeit news. There are two other advanced methods, the Neural Network Machine and the Support Network (SVM), which achieve 99.90 percent accuracy.

III. PROPOSED SYSTEM

The proposed system for detecting deepfakes by using Convolutional Neural Networks (CNNs) to examine the manipulated elements within images. By examining the patterns, textures, and distortions, the neural network can compare in between the original and manipulated media. The system checks the difference between the elements against a dataset of original, unaltered images, making sure that it points out the different characteristics of deepfake presence. This algorithm offers a good level of accuracy and efficiency in identifying the manipulated content, thereby safeguarding against the spread of misleading media content.

Deepfake images are the main objective of ours, which is to be able to accurately identify them in a sea of genuine ones. We have used various techniques and methods to achieve this, using a large dataset of 70,000 images from Kaggle for training our model. In order to achieve optimal performance in differentiating genuine images from deep- fakes, we have fine tuned our CNN based approach through rigorous experimentation and comparative studies.

I

n order to preserve the trust and integrity of digital media, especially in the face of advanced artificial intelligence manipulation, the development of an effective deepfake detection system is of paramount importance. Our system is proactively looking for inconsistencies and patterns that indicate the falsification of content by using advanced algorithms and implementing real-time monitoring mechanisms. With this active approach, users are encouraged to make more effective use of Internet platforms by increasing their awareness and distinguishing between genuine and false content.

Essentially, our proposed system is a digital detective who closely examines the online content to detect what's real and what isn't. We are aiming to strengthen users' confidence in Digital Media and ensure they can rely on the authenticity of information that they come across, through a combination of advanced artificial intelligence technology and efficient detection techniques.

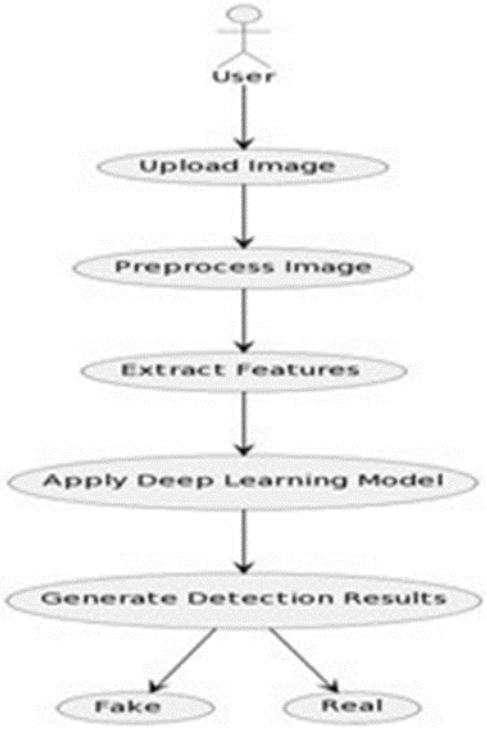

Fig 3.1 Dataflow Diagram

IV. Approach

What is a Convolutional Neural Network (CNN)?

A Convolutional Neural Network (CNN), also known as ConvNet, is a specialized type of deep learning algorithm mainly designed for tasks that necessitate object recognition, including image classification, detection, and segmentation. CNNs are engaged in a diverse set of realistic scenarios, such as autonomous vehicles, security camera systems, and others.The convolutional neural network is made of: Convolutional layers,Rectified Linear Unit (ReLU for short),Pooling layers Fully connected layers.

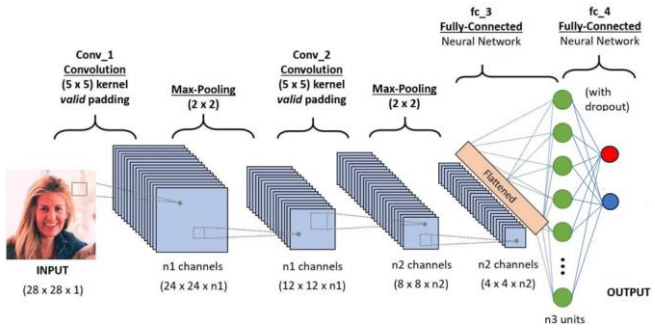

Fig 4.1 CNN

1) Convolution layers: This is the basic component of a CNN. The purposed name suggests, the main mathematical task achieved is called Convolution, which is the application of a sliding window function to a matrix of pixels illustrating an image. The sliding function trained to the matrix is called kernel or filter, and both can be used inversely.
For example, we have an RGB image(4x4x3) that has been separated by its three color planes — Red, Green, and Blue. There are a collection of such color spaces in which images exist — Grayscale, RGB, HSV, CMYK, etc.

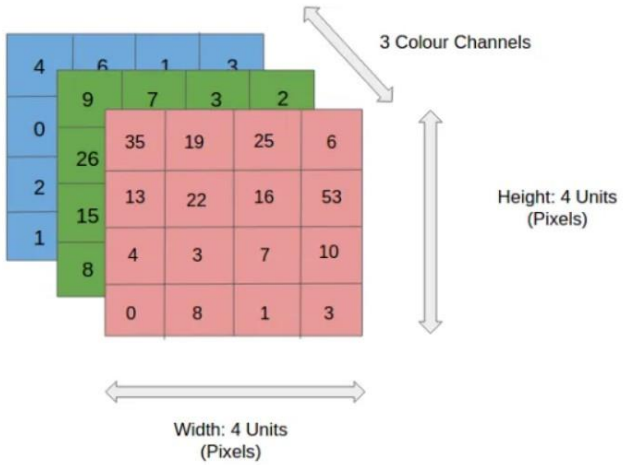

Fig 4.2 RGB Image

Dimensions of Image = 5 (Height) x 5 (Breadth) x 1 (No. of channels)

In the beneath demonstration, the green section is similar to our 5x5x1 input image, I. The element associated in the convolution operation in the initial stage of a Convolutional Layer is called the Kernel/Filter, K, appeared in color yellow. Here we have chosen K as a 3x3x1 matrix.

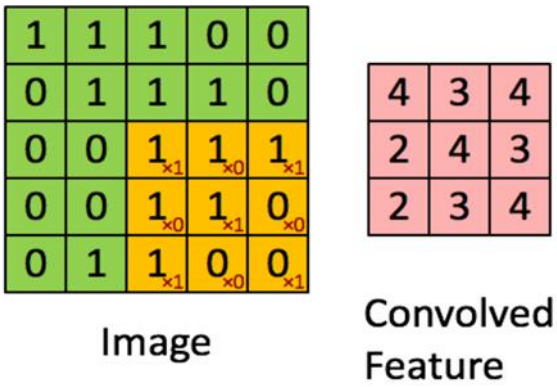

Fig 4.3 Convoluting a 5x5x1 image with a 3x3x1 kernel

2) Activation Function: An activation function, ReLU is applied after every convolution operation. This function support the network to learn non-linear relationships among the features in the image, therefore making the network more powerful for detecting different patterns. In the end, it helps to lighten the vanishing gradient problems.

3) Pooling layer: The ultimate aim of the pooling layer is to pull the most remarkable features from the convoluted matrix. It is possible by seeking few aggregation operations, which decrease the dimension of the feature map (convoluted matrix), hence diminishing the memory used while training the network.

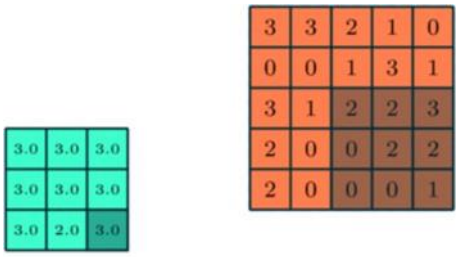

Fig 4.4 3x3 pooling over 5x5 convolved feature

Similar to the Convolutional Layer, the Pooling layer is in charge for minimizing the spatial size of the Convolved Feature. This is to diminish the computational power necessary to process the data through dimensionality reduction. Additionally, it is useful for obtaining dominant features which are rotational and positional invariant, thus conserving the process of productively training the model.

4) Fully connected layers: These are in the final layer of the Convolutional Neural Network, and their inputs correlate to the compact one-dimensional matrix produced by the final pooling layer. ReLU activation functions are applied to them for non-linearity.

Finally, a softmax prediction layer is used to generate probability values for every possible output labels, and the final label anticipated, will be the highest probability score.

V. OUTPUTS

The dataset is taken from Kaggle with the memory size of about 1.8GB that is 70,000 trained and tested data.

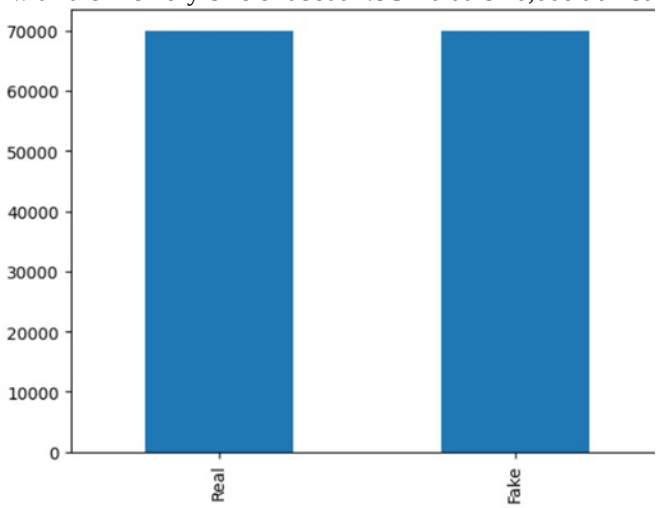

Fig. 5.1 Data Set

The image is the representative of Convolutional Neural Network (CNN) architecture, which is designed for deepfake detection in images. The input layer of network accepts images of shape (75, 75, 3), where 75x75 are the dimensions and 3 represents RGB color channels.

The network here starts with an input layer followed by series of convolutional and max-pooling layers. The first Conv2D layer has 64 filters, each size of (3, 3), and it processes input image to produce an output of shape (75, 75, 64). This specifies that the geometrical dimensions of image are preserved, while depth in changes to 64 due to number of filters.

Next, a MaxPooling2D layer with pool size of (2, 2) lessen the geometrical dimensions by half, resulting an output shape of (37, 37, 64). This process helps in reducing the computational load and capturing dominant features.

The process repeats with another Conv2D layer, but this time with 96 filters of size (3, 3), generating an output of (37, 37, 96). The following MaxPooling2D layer again halves the dimensions, leading to (18, 18, 96).

The third Conv2D layer with 128 filters (3, 3) processes feature maps, producing an output of (18, 18, 128). Another MaxPooling2D layer reduces it to (9, 9, 128).

Finally, a fourth Conv2D layer with 128 filters is applied, which is followed by a MaxPooling2D layer, reducing output to (4, 4, 128). This reduction in geometrical dimensions works while increasing depth allowing the network to capture complex patterns and features similar for deepfake detection.

Each convolutional layer is used to detect the features like edges and textures, while max-pooling layers reduce the dimensions, by making network work more efficient and less vulnerable to overfitting. By the end of these layers, the network transforms the input image into a compact, high-dimensional presentation containing critical features for classification.

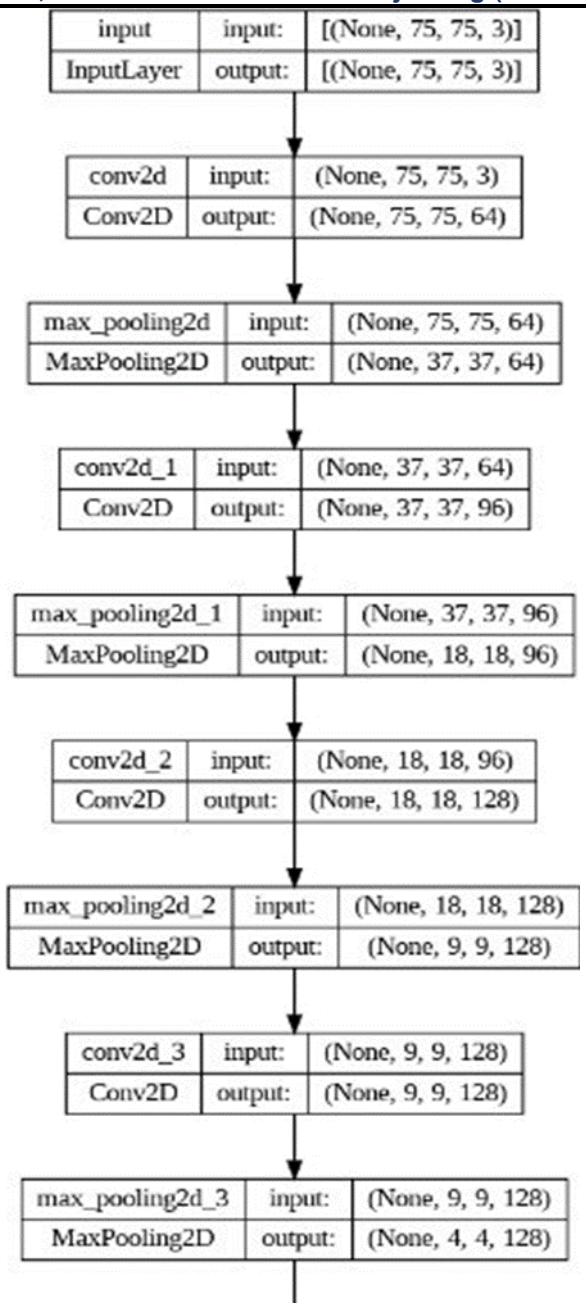

Fig 5.2 The pixel representation of the image that is taken from the dataset is classified taking CNN algorithm.

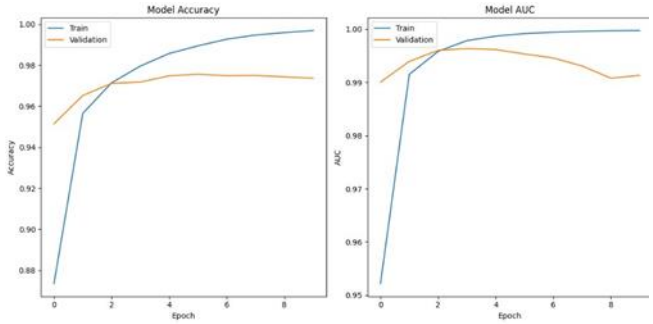

Fig. 5.3 Model accuracy and Model AUC for model1

This is the graph for the taken model validating the Model Accuracy and Model AUC. The model accuracy is the Number of Correct Predictions to the Total Number of Predictions. The Model AUC is used to evaluate the performance of the model.

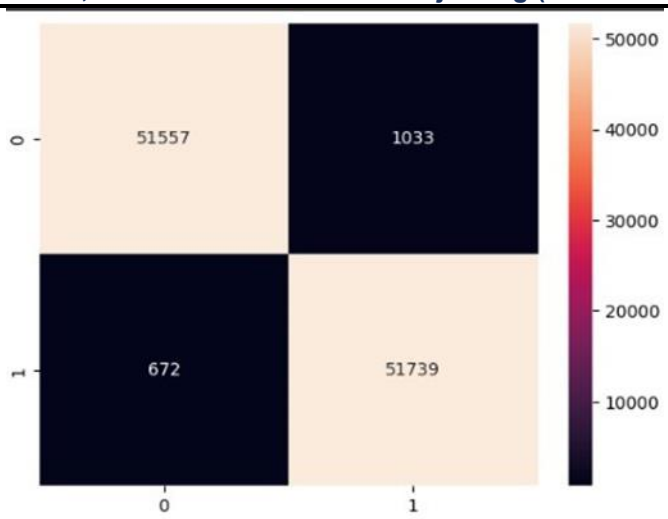

Fig. 5.4 The confusion matrix for the trained model

This is a confusion matrix for a trained model, represented using a heatmap. Here is the explanation of each cell:

- Top left (True Negative - TN): 51,557
- Top right (False Positive - FP): 1,033
- Bottom left (False Negative - FN): 672
- Bottom right (True Positive - TP): 51,739

The confusion matrix provides a detailed explanation of the prediction results on a classification problem, which helps to comprehend the performance of the model by showing how many of the actual labels were accurately projected.

Here are the performance metrics for the trained model based on the given confusion matrix:

- Accuracy: 0.9838 (98.38%)
- Precision: 0.9804 (98.04%)
- Recall: 0.9872 (98.72%)
- F1 Score: 0.9838 (98.38%)

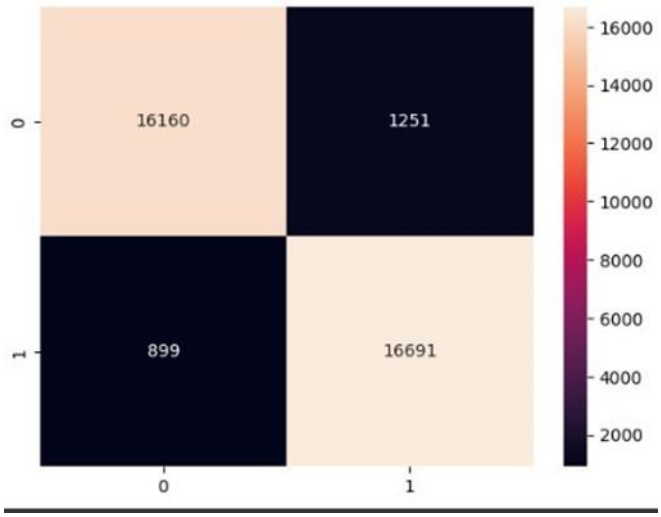

Fig. 5.5 The confusion Matrix for the test model

This is another confusion matrix, this time for a tested model. Let's interpret this matrix as we did previously:

- Top left (True Negative - TN): 16,160
- Top right (False Positive - FP): 1,251
- Bottom left (False Negative - FN): 899
- Bottom right (True Positive - TP): 16,691

Here are the performance metrics for the tested model based on the given confusion matrix:

- Accuracy: 0.9386 (93.86%)
- Precision: 0.9303 (93.03%)
- Recall: 0.9489 (94.89%)
- F1 Score: 0.9395 (93.95%)

The metrics that are indicates above for the tested model also performed well, but with slightly lower values when compared with the trained model's metrics. This submits that the model might not perform good on the new data compared to the training data.

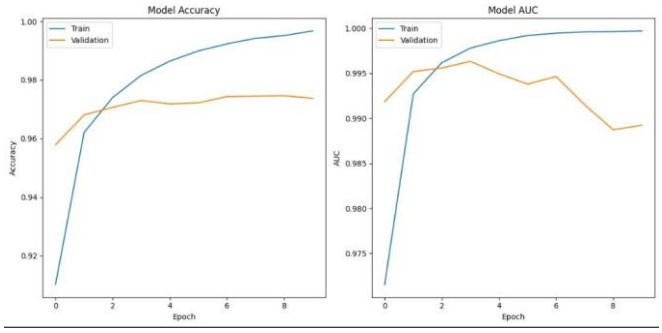

Fig. 5.6 The graph for the VGG19 model validating the Model Accuracy and Model AUC.

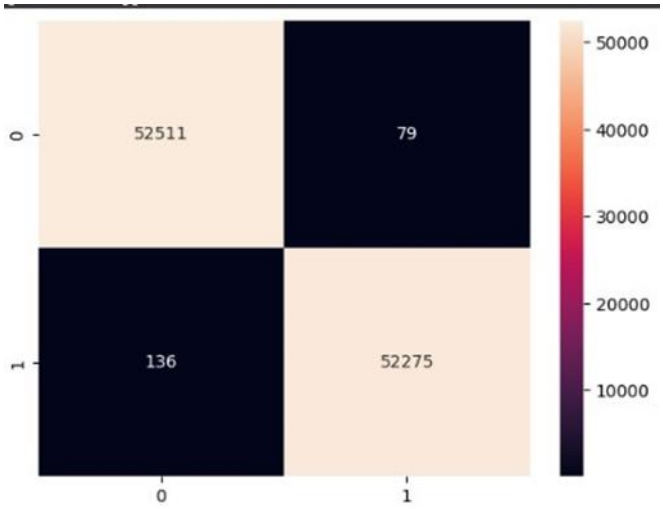

Fig. 5.7 The confusion matrix for trained model VGG18.

This is a confusion matrix for a trained model VGG18, represented using a heatmap. Here is the explanation of each cell:

- True Negative (TN): 52,511
- False Positive (FP): 79
- False Negative (FN): 136
- True Positive (TP): 52,275

Summary

- Accuracy: 99.80%
- Precision: 99.85%
- Recall: 99.74%
- F1 Score: 99.80%

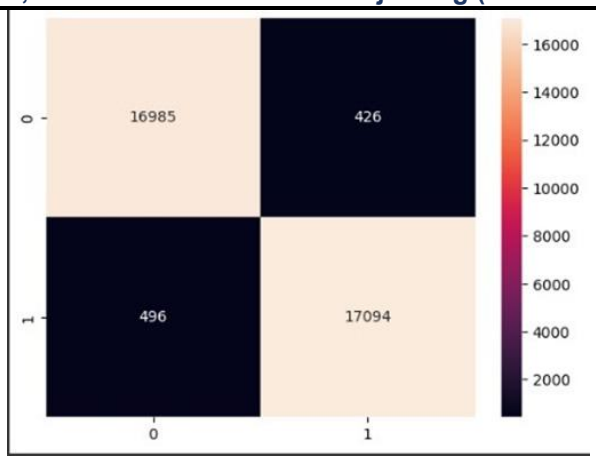

Fig. 5.8 The Confusion matrix for tested model VGG18

Let's interpret this confusion matrix and calculate the relevant performance metrics.

- True Negative (TN): 16,985
- False Positive (FP): 426
- False Negative (FN): 496
- True Positive (TP): 17,094

Summary

- Accuracy: 97.37%
- Precision: 97.57%
- Recall: 97.18%
- F1 Score: 97.38%

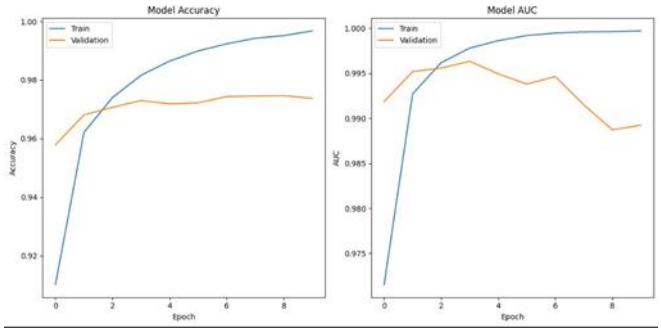

Fig 5.9 The graph for the Mobile Net model validating the model Accuracy and Model AUC.

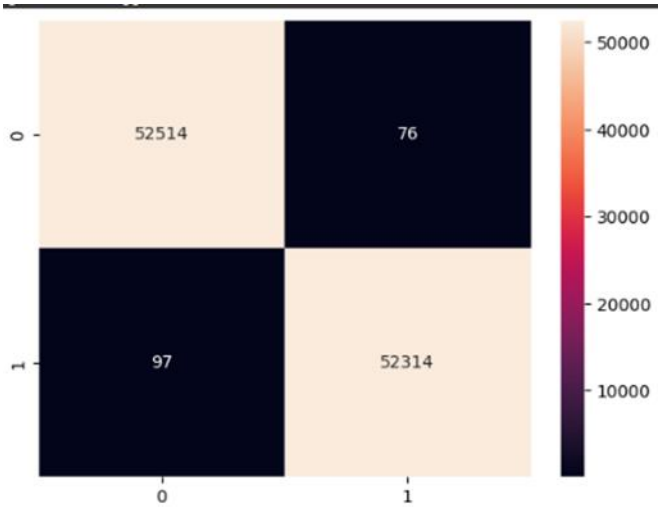

Fig.5.10 The confusion matrix for the trained model Mobile Net.

Let's interpret this new confusion matrix and calculate the relevant performance metrics.

- True Negative (TN): 52,514
- False Positive (FP): 76
- False Negative (FN): 97
- True Positive (TP): 52,314

Summary

- Accuracy: 99.84%
- Precision: 99.85%
- Recall: 99.81%
- F1 Score: 99.83%

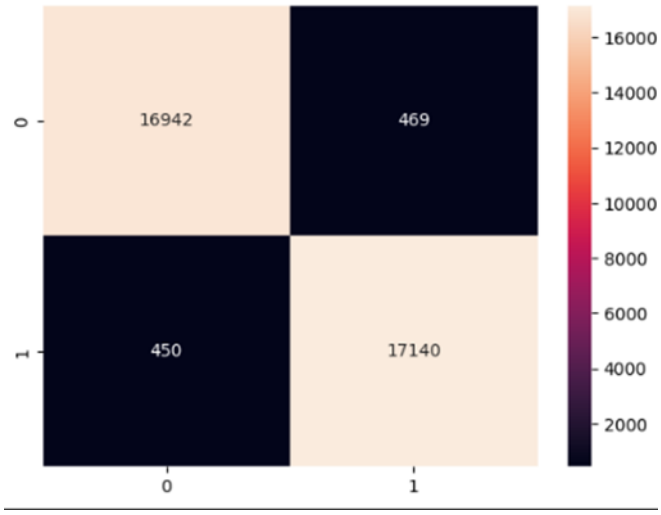

Fig 5.11 The confusion matrix for the tested model Mobile Net.

Here's a breakdown of the values in the confusion matrix:

- **True Positives (TP):** 17140
- **True Negatives (TN):** 16942
- **False Positives (FP):** 469
- **False Negatives (FN):** 450

Summary

- **Accuracy:** 99.07%
- **Precision:** 97.34%
- **Recall:** 97.44%
- **F1 Score:** 97.39%

These metrics suggest that the model is performing quite well in terms of classification accuracy, precision.

VI. CONCLUSION

In conclusion, the proliferation of deepfake technology underscores the pressing need for effective detection methods to preserve the authenticity of digital media. Our project addresses this challenge by harnessing Convolutional Neural Networks (CNNs) to analyze and compare visual elements, enabling accurate identification of manipulated content. Through rigorous research and training on extensive datasets, we have developed a robust approach to distinguish between genuine and deepfake images with precision. By deploying sophisticated algorithms and real-time monitoring systems, we strive to mitigate the spread of deceptive media, ensuring that users can trust the content they encounter online. Our efforts represent a crucial step towards safeguarding the integrity of digital information in the face of evolving AI manipulation techniques, ultimately fostering a more secure and trustworthy digital environment for all.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, 1097–1105.

[2] K. Ravi, (2018). Detecting fake images with Machine Learning. Harkuch Journal

[3] L. Zheng, Y. Yang, J. Zhang, Q. Cui, X. Zhang, Z. Li, et al. (2018). TI-CNN: Convolutional Neural Networks for Fake News Detection. United States.

[4] M. Villan, A. Kuruvilla, K. J. Paul, E. P. Elias, (2017). Fake Image Detection Using Machine Learning. IRACST—International Journal of Computer Science and Information Technology Security (IJCSITS).

[5] S. Aphiwongsophon, P. Chongstitvatana, (2017). Detecting Fake News with Machine Learning Method. Chulalongkorn University, Depart- ment of Computer Engineering, Bangkok, Thailand.

[6] Y. Li, S. Cha, (2019). Face Recognition System. arXiv preprint arXiv:1901.02452.

[7] R. Saracco, (2018). Detecting fake images using artificial intelligence. IEEE Future Directions

[8] Akhtar, Z., Mouree, M. R., Dasgupta, D. (2020). Utility of deep learning features for facial attributes manipulation detection. Paper presented at the 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI).

[9] Bekci, B., Akhtar, Z., Ekenel, H. K. (2020). Cross-dataset face manipulation detection. Paper presented at the 2020 28th Signal Processing and Communications Applications Conference (SIU).

[10] Zhang, L.; Qiao, T.; Xu, M.; Zheng, N.; Xie, S. Unsupervised learning- based framework for deepfake video detection. IEEE Trans. Multimed. 2022, 25, 4785–4799.

[11] Ju, Y.; Jia, S.; Cai, J.; Guan, H.; Lyu, S. GLFF: Global and Local Feature Fusion for Face Forgery Detection. arXiv 2022, arXiv:2211.08615

[12] G. Patrini, F. Cavalli and H. Ajder, ”The state of deepfakes: Reality under attack”, 2018

[13] F. Matern, C. Riess and M. Stamminger, ”Exploiting visual artifacts to expose deepfakes and face manipulations”, Proc. IEEE Winter Appl. Compute Vis. Workshops (WACVW), pp. 83-92, Jan. 2019.

[14] L. Guarnera, O. Giudice and S. Battiato, ”Fighting deepfake by ex- posing the convolutional traces on images”, arXiv:2008.04095, 2020.

[15] Symeon, P.C.G.K.Z.; Kompatsiaris, P.I. AFace PREPROCESSING APPROACH FOR IMPROVED DEEPFAKE DETECTION. arXiv 2020, arXiv:2006.07084

[16] K. Songsri-in and S. Zafeiriou, ”Complement face forensic detection and localization with faciallandmarks”, arXiv:1910.05455, 2019.

[17] R. Durall, M. Keuper, F.-J. Pfreundt and J. Keuper, ”Unmasking DeepFakes with simple features”, arXiv:1911.00686, 2019.