DETECTION AND LOCALISATION OF DEEPFAKES USING DEEP LEARNING

Mohan Aravind Bhavanam[1], Jasvanth Kumar Pallapothu[2], Grandhi Venkata Naga Sai Sasank[3], MD Abdul Razzaq[4]

[1]UG Student, Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India
[2]UG Student, Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India
[3]UG Student, Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India
[4]UG Student, Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India

ABSTRACT

In today's digital era, the proliferation of deepfake technology has raised significant concerns about the authenticity and integrity of multimedia content. Deepfake involves the creation of highly realistic fake images and videos and misuse for spreading fake news, defaming individuals, and posing a significant threat to the integrity of digital content. To combat this challenge, our project "Deep learning approaches for robust deepfake detection and localization "aims to address this critical issue by developing a robust system for the identification and localization of deepfake content by using the DenseNet121 model. Densenet's dense connections contribute to improved gradient flow, enabling better information propagation through the network leading to better classification results. The accurate identification of tampered regions holds the utmost importance for identifying the intentions of the offenders. For localization purposes, we use binary masking by segmenting the image with a specific threshold value and indicating the morphed regions. This proposed framework seamlessly integrates forgery detection and localization.

Keywords: *- Deepfake, DenseNet121, Binary mask*

1. INTRODUCTION

In recent years, the proliferation of deepfake technology has raised significant concerns regarding its potential misuse in manipulating media content for malicious purposes. Deepfakes, generated using deep learning algorithms, can produce highly realistic synthetic images, videos, or audio recordings, often indistinguishable from genuine ones. As a result, they pose a serious threat to various aspects of society, including national security, politics, and individual privacy. The ability to detect and localize deepfake content has become a critical area of research and development. Effective detection and localization methods are essential for mitigating the harmful effects of deepfakes and safeguarding the integrity of multimedia content. By accurately identifying and isolating manipulated elements within media files, it becomes possible to prevent the dissemination of misinformation, protect individuals from targeted attacks, and preserve trust in digital media platforms.

This project aims to contribute to the ongoing efforts in deepfake detection and localization by proposing novel algorithms and techniques. Leveraging advancements in machine learning, computer vision, and signal processing, we seek to enhance the accuracy and efficiency of existing detection methods. Our approach focuses on analyzing subtle discrepancies in facial expressions, voice characteristics, and contextual cues to distinguish between genuine and deepfake content. Furthermore, the project emphasizes the importance of real-time detection and localization capabilities, considering the rapid spread of deepfake content across online platforms. By developing algorithms that can quickly identify and pinpoint manipulated regions within multimedia files, we aim to empower users with the tools needed to verify the authenticity of digital content in a timely manner. Ultimately, the successful implementation of this project will contribute to the advancement of deepfake detection and localization technology, bolstering defenses against the proliferation of synthetic media manipulation. Through collaborative efforts with researchers, policymakers, and industry stakeholders, we aspire to foster a safer and more trustworthy digital environment for all users.

2. RELATED WORK

2.1 Deepfake detection

Most approaches to face forgery detection approach it as a binary (real/fake) classification problem. Many deep learning-based techniques have been developed to identify facial forgeries. Different blink frequencies were noted by Li et al. [1] between movies that were faked and those that were real. The classification of video authenticity was made possible by the authors' use of CNNs and Long Short-Term Memory (LSTM) [2] models to extract blink-related information in the temporal domain. An attention mechanism was added by Dang et al. [3] to highlight the forgery area, which increased the accuracy of forgery classification. As an alternative, Nguyen et al. [4] suggested using a capsule network that was created especially to recognize fake photos or videos. In order to improve the model's overall performance, Fang et al. [5] also added the integration of reconstruction losses in addition to classification.

2.2 Combined Detection and Localization

Multi-task learning was suggested by Nguyen et al. [5] as a way to identify and locate altered areas in both photos and movies. By emphasizing the forged regions, attention processes [6,7] improve the feature maps of the classification job. The detection of edge regions with heterogeneous boundaries between the altered face and background was suggested by Li et al. [8]. Nevertheless, no research has yet been done on vision segmentation foundation models for simultaneous face forgery detection and localization.

2.3 Existing model overview

Currently, the most of the existing forgery detection methods utilize traditional convolutional neural networks (CNNs) for feature extraction and classification which classifies whether the image is real or fake. Despite continuous advancements in forged face detection technology in recent years, accurate localization of forged regions remains a challenge, particularly for models that solely provide classification results. The limitations of the existing system based on the convolutional neural network include limited temporal understanding, computational requirements, and training data imbalance. Segment Anything Model (SAM) [9] is used in the existing model for localization purposes. However, the results of segmentation do not clearly visualize the morphed regions.

3. PROPOSED MODEL

DenseNet121 model is used for classifying whether an image is real or fake and Binary masking for Localization of forged regions. DenseNet's dense connectivity patterns facilitate improved gradient flow during training. This helps in combating the vanishing gradient problem and enables better information propagation through the network. As a result, the network can learn more complex features effectively. With dense connections, each layer in DenseNet receives feature maps from all preceding layers. DenseNet's architecture tends to perform well even with limited training data. This is because of its ability to efficiently utilize available data and extract relevant features effectively. Incorporating binary masking for localization purposes allows for accurate identification of morphed regions in faces. By computing the difference in intensity values of pixels beyond a threshold, the system can effectively isolate regions of interest, aiding in the detection of fake faces. The primary objective of the proposed system is to achieve high accuracy in distinguishing between fake and real faces while minimizing false positives and false negatives.

3.1 Architecture of the proposed model

A cutting-edge architecture called DenseNet121 has demonstrated exceptional performance in image classification challenges. DenseNet, which stands for "Densely Connected Convolutional Networks," gets its name from the feedforward connections it makes between each layer and every other layer. Two salient characteristics of DenseNet distinguish it from other CNN architectures. Its dense block structure is the first feature; every layer is feed forwardly coupled to every other layer. Secondly, it makes use of bottleneck layers, which assist in lowering the number of parameters without lowering the total amount of characteristics the network learns. Every convolutional layer of a conventional feed-forward convolutional neural network (CNN), with the exception of the first, which receives input, gets the output of the preceding convolutional layer and generates an output feature map, which is then forwarded to the following convolutional layer. As a result, there are 'L' direct connections for 'L' layers, one for each layer and the layer after that. The 'vanishing gradient' issue, however, appears as the CNN gets deeper—that is, as the number of layers increases. This implies that certain information may vanish or get lost when the path for information from the input to the output layers lengthens, which lowers the network's capacity for efficient training. By altering the conventional CNN architecture and streamlining the layer-to-layer connectivity structure, DenseNets alleviate this issue. Densely Connected Convolutional Network is the name given to an architecture in which every layer is directly connected to every other layer: DenseNet. There are L(L+1)/2 direct connections for 'L' layers.

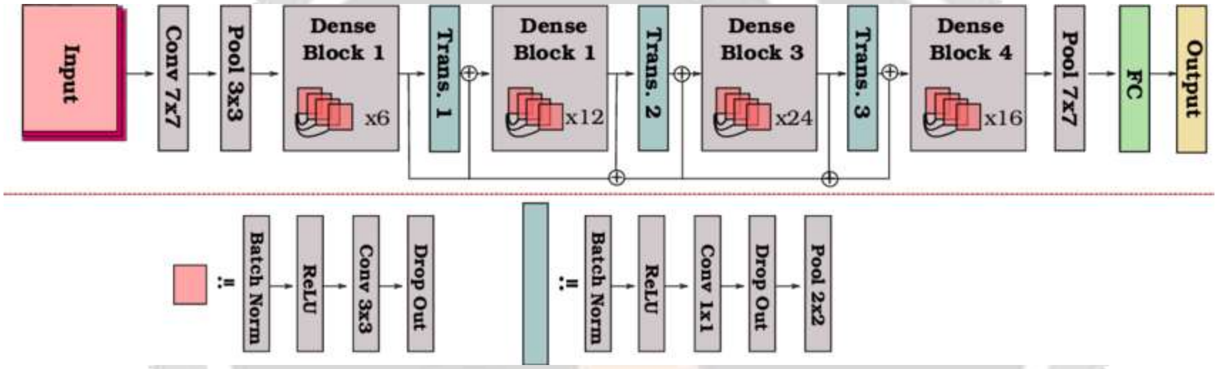


Fig 1: DenseNet121 Architecture

3.2 Working of the model

The Deepfake detection and localization system using DenseNet121 and binary masking operates through a sequence of steps to accurately classify and localize the deepfakes. Firstly, the system is given a dataset consisting of a combination of real and fake images obtained from publicly available repositories like Kaggle. These images undergo preprocessing step which includes image resizing, noise removal, and pixel value normalization thereby standardizing the data. Then, the processed dataset is fed to DenseNet121 model which learns features from the input images. As the images pass through the DenseNet121 layers, low level features like edges, lines, and corners are identified in early layers and high-level features like shapes, textures and local object parts are more focused in the deeper layers. The training phase unfolds iteratively, with batches of preprocessed images being fed into the model. Throughout this iterative process, the model fine-tunes its parameters to minimize the disparity between predicted and actual labels, thereby refining its ability to distinguish between authentic and manipulated imagery. Upon completion of training, the efficacy of the trained model is scrutinized through validation and testing using a distinct dataset. A set of evaluation metrics, including accuracy, precision, recall, and F1 score, is computed to gauge the model's performance in deepfake detection and localization. ‘140k Real and Fake Faces’ is the dataset used for training the DenseNet121 model which contains images of various sizes so in the preprocessing step all the images are resizes to 224*224 pixels. The localization is carried out using Binary masking. If the image is identified as fake image, then the image is first converted to gray scale and then a specific threshold is selected to convert that gray scale image to binary image. Regions highlighted in white within the binary image signify areas where tampering has occurred, facilitating precise localization of deepfake alterations.

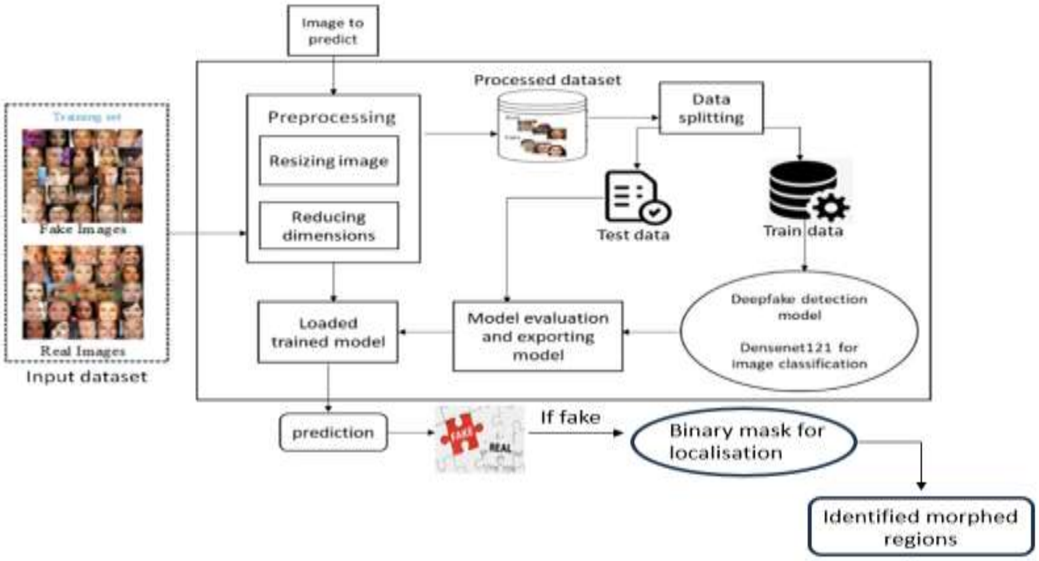


Fig 2: Process flow of proposed system

3.3 Dataset Analysis

The dataset utilized in this project, named '140k Real and Fake Faces,' comprises 70,000 authentic facial images sourced from the Flickr dataset curated by Nvidia, alongside 70,000 fabricated facial images sampled from the extensive 1 million FAKE faces corpus produced by StyleGAN. This amalgamation ensures a diverse representation of both real-world and synthetic facial features. The dataset is meticulously partitioned into training, testing, and validation subsets to facilitate robust model development and evaluation. Each subset maintains a balanced distribution of real and fake faces, with an equal 50% split between the two categories, fostering fair learning conditions.

Within this framework, the 140,000 images are meticulously allocated, with 20,000 images dedicated to validation, evenly distributed between 10,000 real and 10,000 fake faces. The training set encompasses 100,000 images, consisting of 50,000 real faces and an equivalent number of fake faces, ensuring parity in class representation during model training. Lastly, the testing set comprises 20,000 images, meticulously divided into 10,000 real and 10,000 fake faces, serving as an independent benchmark to assess model performance on unseen data. This comprehensive dataset structure, meticulously organized and balanced, underpins the efficacy and reliability of the ensuing model for discerning real from fake facial images.

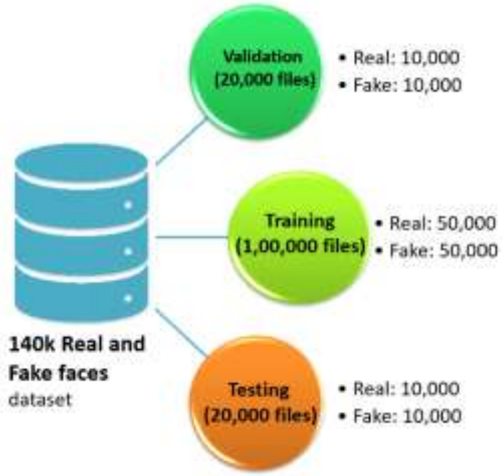


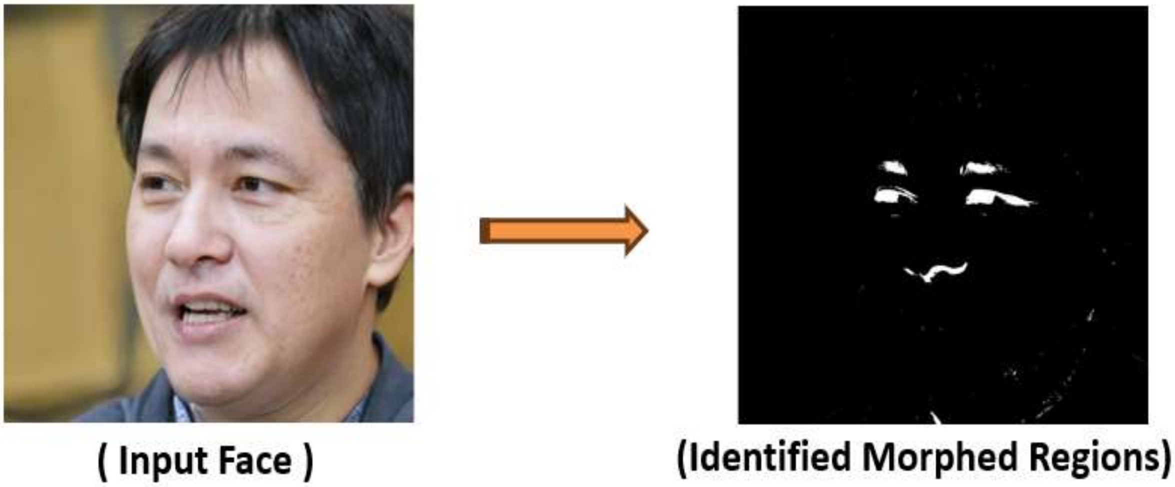
Fig 3: Dataset description

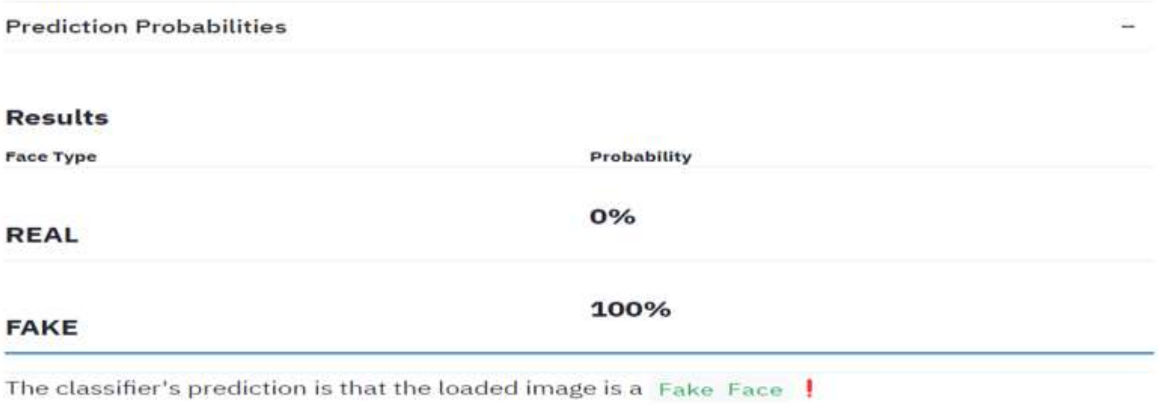
4. RESULTS AND PERFORMANCE ANALYSIS

4.1 Training data

The training process entails the iterative refinement of a neural network model designed to distinguish between real and fake faces. Initially, the dataset of 140,000 images is prepared, involving standard preprocessing steps and partitioning into training, validation, and testing subsets. The model architecture, likely comprising convolutional and fully connected layers, is then defined using Keras, and the model is compiled with specific optimization settings and evaluation metrics. Training unfolds over multiple epochs, with each epoch involving the sequential processing of batches of training data through the network. The model updates its parameters based on computed loss, aiming to minimize the difference between predicted and actual labels. Simultaneously, validation datasets are used to monitor generalization and detect overfitting. Progress logs track metrics such as loss and accuracy for both training and validation datasets, with a focus on improving validation loss over epochs. Upon completion of training, the final model is evaluated on a separate testing dataset to assess its generalization performance. The process continues until a stopping criterion is met, such as convergence or a predefined number of epochs. Overall, this training process exemplifies the iterative refinement of a neural network model to effectively classify images as real or fake, demonstrating the robustness and adaptability of deep learning techniques in image recognition tasks**.**

4.2 Model prediction results





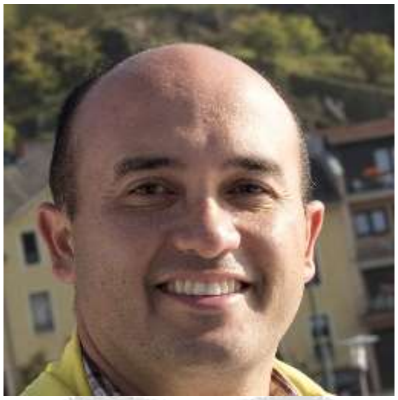
Prediction Probabilities –

Results

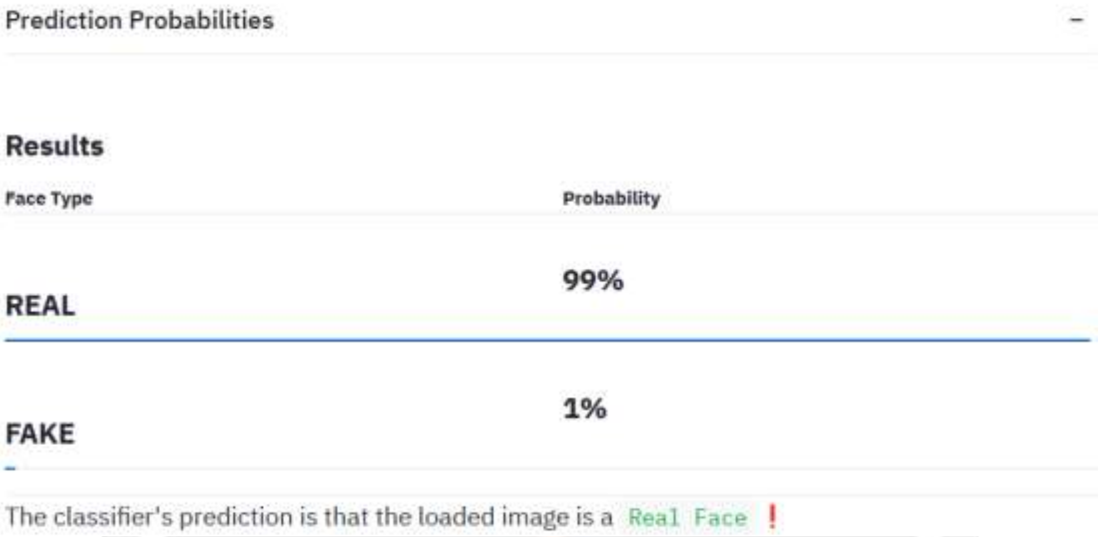
Face Type	Probability
REAL	0%
FAKE	100%

The classifier's prediction is that the loaded image is a Fake Face !

Fig 4: Prediction result for sample input 1



(Input face)



Prediction Probabilities

Results

Face Type	Probability
REAL	99%
FAKE	1%

The classifier's prediction is that the loaded image is a `Real Face` !

Fig 5: Prediction result for sample input 2

4.3 Loss and Accuracy plots

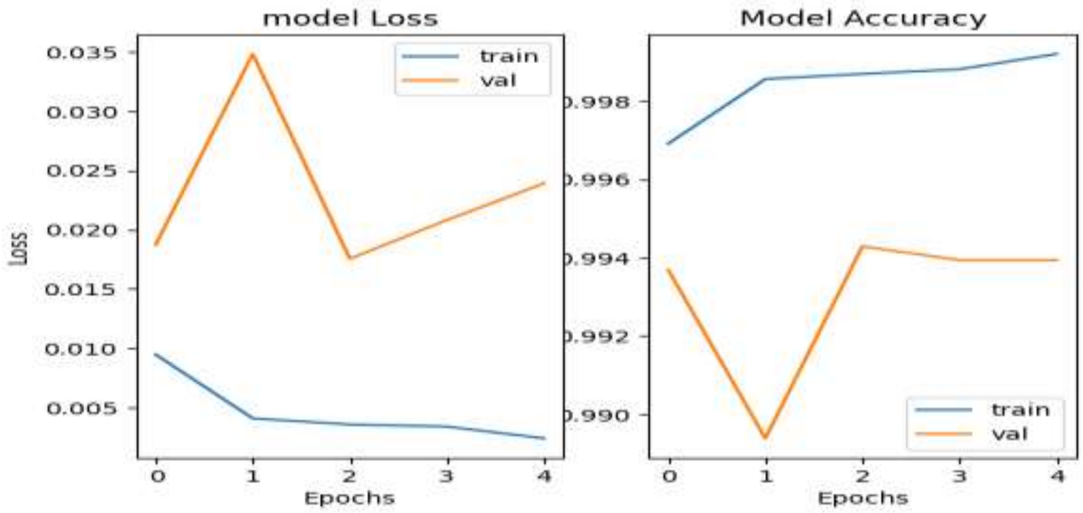


Fig 6: (a) Loss vs epochs (b) Accuracy vs epochs

Both graphs shown above depict how these metrics change over a number of epochs, which are iterations of the training process for a machine learning model. The model loss graph shows a downward trend, which means the model's loss is decreasing as the number of epochs increases. This suggests the model is learning and improving its performance over time. The model accuracy graph also shows an upward trend, which means the model's accuracy is increasing as the number of epochs increases. This is consistent with the decreasing loss since a lower loss typically indicates better accuracy. The y-axis for model accuracy is labeled, and it shows that the model's accuracy is very high, around 99.43 % at the end of the training process. Overall, the graphs suggest that the machine learning model is performing well and has learned to accurately classify the data it was trained on.

4.4 Confusion Matrix

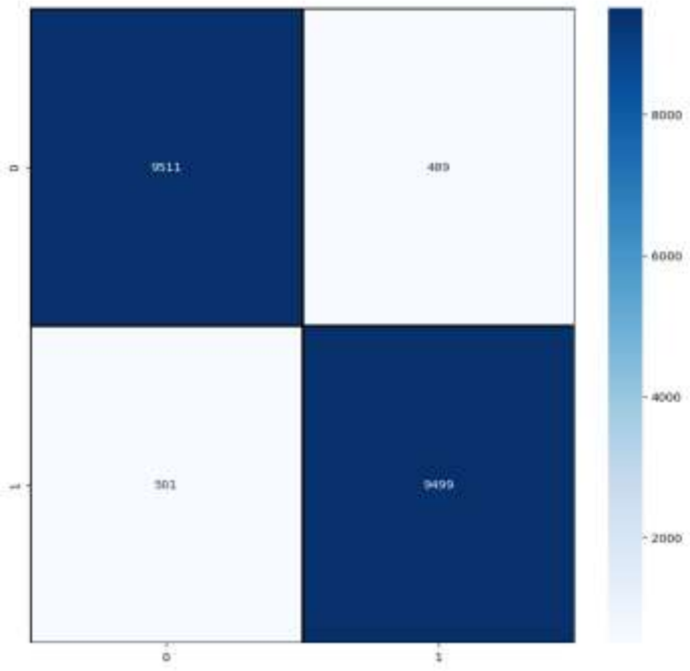
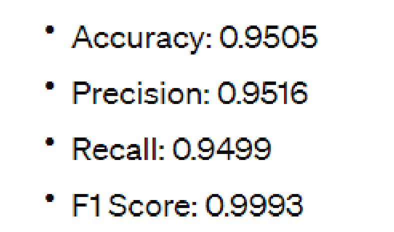


Fig 7: Confusion matrix

The obtained confusion matrix is [[9511 489
501 9499]]



- Accuracy: 0.9505
- Precision: 0.9516
- Recall: 0.9499
- F1 Score: 0.9993

Fig 7: Evaluation Metrics

The high counts of true positives and true negatives indicate that the model is effective at correctly identifying instances from both classes. This suggests that the model has a strong ability to discriminate between positive and negative cases. The relatively low counts of false positives and false negatives indicate that the model's errors are minimal. This implies that the model maintains a good balance between sensitivity (the ability to correctly identify positive cases) and specificity (the ability to correctly identify negative cases). Based on the confusion matrix, the model appears to perform well in terms of both accuracy and reliability. However, it's essential to consider additional performance metrics (such as precision, recall, and F1 score) to gain a comprehensive understanding of the model's performance across different evaluation criteria. In summary, the detailed analysis of the confusion matrix provides valuable insights into the model's classification performance, allowing for informed decisions regarding potential optimizations or enhancements.

5. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The results of successful deepfake detection and localization extends far beyond mere expectation. This proposed model using DenseNet121 architecture and binary masking has yielded promising results., exhibiting a remarkable accuracy of 99.43%. Through extensive experimentation and analysis, we have demonstrated the efficacy of our approach in accurately identifying deepfake content and localizing manipulated regions within multimedia files. The DenseNet121 model exhibited robust performance in classifying the images as genuine or manipulated, achieving high accuracy rates and effectively distinguishing between authentic and deepfake content. Furthermore, the incorporation of binary masking techniques enabled precise localization of manipulated regions within media files, providing valuable insights into the areas affected by synthetic alterations. By identifying and isolating these regions, our method facilitates the mitigation of potential risks associated with deepfake dissemination, including misinformation, privacy breaches, and targeted attacks. Overall, our project contributes to the ongoing efforts in combating the spread of deepfake content and safeguarding the integrity of digital media platforms. The combination of DenseNet121 classification and binary masking localization offers a promising approach to detecting and mitigating the impact of synthetic media manipulation, empowering users with the tools needed to verify the authenticity of multimedia content and preserve trust in digital communication channels. In conclusion, our project represents a significant step towards addressing the challenges posed by deepfake technology, emphasizing the importance of collaborative efforts in developing innovative solutions to protect the integrity and authenticity of multimedia content in the digital age.

5.2 Future Scope

Future research directions may focus on enhancing the scalability, efficiency, and robustness of deepfake detection and localization systems. Additionally, exploring the integration of multimodal analysis techniques and advanced adversarial detection methods could further strengthen the resilience of anti-deepfake solutions in combating evolving threats in the digital landscape. One potential direction is the integration of multi-modal analysis techniques to augment the capabilities of the detection system. By incorporating additional modalities such as audio and text, it may be possible to create more comprehensive models capable of detecting sophisticated deepfake content across various media formats. This could enhance the robustness of the system and improve its performance in identifying manipulated content that may evade traditional visual-based detection methods. Moreover, ongoing advancements in deep learning architectures and algorithms offer opportunities to improve the performance and efficiency of the detection system further. Continual refinement and optimization of the DenseNet121 model, as well as exploration of novel architectures tailored specifically for deepfake detection, could lead to significant advancements in the field. Additionally, leveraging techniques such as transfer learning and domain adaptation could enable the adaptation of pre-trained models to new domains and scenarios, enhancing the versatility and applicability of the detection system across different contexts. In conclusion, the project on deepfake detection and localization presents a rich landscape for future research and development, with numerous opportunities for innovation and advancement. By exploring these avenues, researchers can contribute to the ongoing efforts to combat the proliferation of deepfake content and safeguard the integrity of digital media platforms in an increasingly complex and dynamic landscape.

6. REFERENCES

[1]. LIY, C.M., InIctuOculi, L.: Exposingaicreated fakevideosbydetectingeyeblinking. In: IEEE WIFS (2018)
[2]. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation (1997)
[3]. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K. On the detection of digital face manipulation. In: IEEE CVPR (2020)
[4]. Nguyen, H.H., Yamagishi, J., Echizen, I.: Capsule-forensics: Using capsule networks to detect forged images and videos. In: IEEE ICASSP (2019)

[5]. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. In: IEEE BTAS (2019)

[6]. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: IEEE CVPR (2020)

[7]. Kong, C., Chen, B., Li, H., Wang, S., Rocha, A., Kwong, S.: Detect and locate: Exposing face manipulation by semantic-and noise-level telltales. IEEE TIFS (2022)

[8]. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: IEEE CVPR (2020)

[9]. arXiv:2306.17075 [cs.CV]

[10]. Budati, M., Karumuri, R. An intelligent lung nodule segmentation framework for early detection of lung cancer using an optimized deep neural system. Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023-17791-8