iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70706

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Deep Fake Using Audio and Image Detection

Dr. S. Udayashree[1], Dr. R.G. Suresh Kumar[2], Ms. S. Tharunika[3], Ms. P. Ramadevi[4], Ms. D. Latika[5], Ms. S. Pavithra[6]
[1, 2]*Professor, RGCET, Puducherry*
[3, 4, 5, 6]*B.TECH (CSE), RGCET, Puducherry*

Abstract: ***In the AI-driven era, deep fakes, generated through advanced techniques like Generative Adversarial Networks (GANs), present significant threats by creating highly realistic yet fabricated media. While audio deep fakes have received considerable attention, the detection of manipulated images remains underexplored, creating a critical gap in comprehensive deep fake identification. Our proposed system bridges this gap by integrating transfer learning for enhanced detection across both fake audio and manipulated images. For image analysis, we utilize the VGG19 architecture, leveraging its deep convolutional layers and pre-trained weights to effectively identify visual artifacts and manipulations. For audio detection, we employ a hybrid model combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), where the CNN extracts spatial features from audio spectrograms, while the RNN captures temporal dependencies to ensure robust analysis of audio authenticity. This hybrid approach allows for a comprehensive, dual-modal detection system that addresses both visual and auditory deep fakes. By combining these methodologies, the system ensures no manipulation indicators are overlooked, providing enhanced reliability and security in detecting digital content tampering. Ultimately, our system contributes to safeguarding the integrity of information, offering a powerful tool to combat the evolving threat of deep fakes in the digital landscape.***

Keywords: Multimedia forensics, Recurrent Neural Networks (RNN), web application.

I. INTRODUCTION

Fake audio and images have become increasingly sophisticated due to the advancements in deep learning techniques, particularly Convolutional Neural Networks (CNNs). These technologies enable the creation of synthetic media that closely resembles authentic sounds or visuals, making it challenging to differentiate between genuine and fabricated content. This capability has given rise to deepfakes, where audio and visual elements are altered or completely fabricated to impersonate individuals or misrepresent events. The ability to generate realistic fake media poses significant ethical and legal concerns, as it can be exploited for malicious purposes, such as spreading misinformation or damaging reputations.

The implications of fake media extend beyond personal identity theft; they can influence public opinion, disrupt political processes, and undermine trust in digital communication. For instance, deepfake technology can create realistic videos of public figures saying or doing things they never actually did, leading to public confusion and manipulation. This potential for abuse raises urgent questions about the authenticity of the information consumed in the digital age. As such, detecting and preventing the misuse of fake media has become a critical challenge for digital security and authenticity verification.

To address these challenges, researchers and technologists are actively developing detection methods to identify manipulations in audio and visual content. Techniques such as analyzing inconsistencies in pixel patterns, audio waveforms, and metadata are being explored to unveil deceptive media. Additionally, there is a growing need for robust frameworks and policies to regulate the use of synthetic media, ensuring that technology serves the public good rather than enabling deception. As the landscape of digital media continues to evolve, ongoing efforts in detection and policy-making will be vital in maintaining trust and integrity in communication.

II. LITERATURE SURVEY

1) *Deepfake detection in digital media forensics [1]* This paper introduces a novel method for detecting Deepfake videos by integrating ResNext, a powerful Convolutional Neural Network (CNN) adept at extracting complex image features, with Long Short-Term Memory (LSTM) networks, which are designed for analyzing temporal sequences in video data. The combination allows for a more comprehensive analysis, as ResNext captures spatial features from individual frames while LSTM processes the temporal dynamics across frames. This synergistic approach enhances the model's ability to identify subtle manipulations in Deepfake videos, improving detection accuracy and robustness against increasingly sophisticated synthetic media.

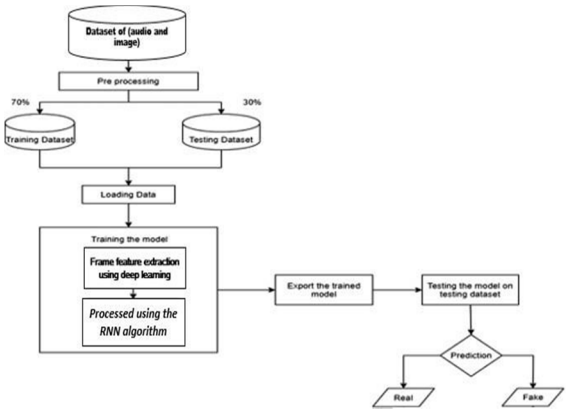
2) *Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence.* This study explores a range of automated techniques for both detecting and generating deepfakes in audio and images. It provides a comprehensive review of existing frameworks, algorithms, and tools designed to identify synthetic media, highlighting their effectiveness and limitations. Additionally, the study examines the application of these methods in various contexts, such as social media, journalism, and security, to address the growing challenge of disinformation. By analyzing the current landscape of deepfake detection technologies, the research aims to identify best practices and potential strategies for mitigating the risks associated with manipulated content in digital environments.
3) *Conspiracy thinking and social media use are associated with ability to detect deepfakes [3]* This study explores a range of automated techniques for both detecting and generating deepfakes in audio and images. It provides a comprehensive review of existing frameworks, algorithms, and tools designed to identify synthetic media, highlighting their effectiveness and limitations. Additionally, the study examines the application of these methods in various contexts, such as social media, journalism, and security, to address the growing challenge of disinformation. By analyzing the current landscape of deepfake detection technologies, the research aims to identify best practices and potential strategies for mitigating the risks associated with manipulated content in digital environments.
4) *Audio-deepfake Detection: Adversarial attacks and countermeasures [4]* The study highlights a significant vulnerability in state-of-the-art audio deepfake classifiers, including the Deep4SNet model, revealing that these systems are susceptible to adversarial attacks. Adversarial attacks involve introducing subtle perturbations to the input data, which can lead classifiers to misidentify manipulated audio as genuine. This vulnerability poses a serious threat to the reliability of audio deepfake detection technologies, as even minor alterations can drastically reduce the accuracy of these classifiers. The findings underscore the necessity for ongoing research and development in robust detection methods that can withstand such attacks, ensuring the integrity of audio verification processes in an increasingly deceptive digital landscape.
5) *Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes [5]* Hyper-realistic deepfakes that incorporate believable content manipulation are often perceived as more credible by viewers, leading them to accept the fabricated media as authentic. This heightened realism creates a sense of trust, making it easier for audiences to be misled by the manipulated content. Conversely, deepfakes featuring less plausible content manipulation can paradoxically have a more significant impact on undermining the legitimacy of the depicted politician. When viewers recognize discrepancies or implausibilities, even in a less realistic context, it can erode trust in the individual and their message, ultimately damaging their reputation and credibility. This dynamic illustrates the complex interplay between the perceived realism of deepfakes and their potential to influence public perception, emphasizing the need for critical media literacy in the age of synthetic media.

III. PROPOSED SYSTEM

The proposed system employs a hybrid approach that utilizes Recurrent Neural Networks (RNN) to effectively detect manipulations in both audio and image content.

By accommodating the temporal dynamics inherent in audio data and integrating advanced image processing capabilities, this model enhances the accuracy of manipulation detection. The system trains audio and image datasets separately, allowing for specialized training that improves detection reliability for each media type. A user-friendly web application, developed using Bootstrap, offers an intuitive interface for users to upload and analyze multimedia content seamlessly. By incorporating RNN, the system significantly increases accuracy in identifying fake audio and manipulated images, thereby enhancing the overall effectiveness of multimedia forensics. This comprehensive approach not only expands the detection scope but also lays a solid foundation for future research, promoting the ongoing advancement of techniques for identifying multimedia manipulations in an increasingly complex digital environment.

A. *Architecture Of The Proposed System:*



The diagram shows the process for training and testing a deep learning model to detect real versus fake content, such as deepfakes, using audio and image data. The dataset is split into 70% for training and 30% for testing. During the training phase, deep learning techniques are used for frame-level feature extraction. These features are then processed by a Recurrent Neural Network (RNN), which is well- suited for handling sequential data. After training, the model is exported and tested on the remaining 30% of the dataset. The testing phase evaluates the model's ability to classify the input data, predicting whether it is "Real" or "Fake." This workflow leverages RNNs for their strength in processing sequences, making it effective for identifying manipulations in multimedia data such as deepfake detection, where both audio and image cues are essential for accurate classification.

IV. IMPLEMENTATION DETAILS

A. Data Collection

The data used in this process is collected from Kaggle, an opensource platform that provides a variety of datasets for machine learning tasks. Kaggle offers high-quality datasets, including those for audio and image data, which are commonly used for training deep learning models. In this case, the dataset likely contains multimedia content, enabling the model to be trained to detect real versus fake (e.g., deepfakes). Using Kaggle's open-source data ensures access to a large and diverse dataset, crucial for building an accurate and robust model. https://www.kaggle.com/datasets/mohammedabdel dayem/the-fake-or-real-dataset - fake audio https://www.kaggle.com/datasets/awsaf49/artifact- dataset - fake image

B. Pre-processing:

Pre-processing plays a crucial role in preparing both the audio and image datasets for effective manipulation detection. For the audio data, we employ Mel-Frequency Cepstral Coefficients (MFCC), a widely used feature extraction technique that captures the spectral characteristics of audio signals. MFCC transforms the audio waveform into a more compact representation, highlighting important features while reducing noise, which enhances the model's ability to distinguish between genuine and manipulated audio. On the other hand, the image dataset is processed separately through a series of techniques that include normalization, resizing, and augmentation. Normalization ensures consistent pixel value ranges, while resizing adjusts images to a uniform dimension for batch processing. Data augmentation introduces variations such as rotation, flipping, and scaling to increase the diversity of training samples, helping the model generalize better to unseen data. This comprehensive pre-processing approach ensures that both audio and image data are optimally prepared, ultimately improving the accuracy and reliability of the manipulation detection system.

C. Feature Extraction

Feature extraction plays a vital role in detecting audio and image manipulations. For audio data, Mel-Frequency Cepstral Coefficients (MFCC) are used to transform the audio waveform into a compact representation that emphasizes essential frequency components. This method captures the perceptually meaningful characteristics of sound, enabling the model to effectively distinguish between genuine and manipulated audio. In the case of images, Recurrent Neural Networks (RNNs) are employed to automatically learn and extract important visual patterns over time, capturing temporal dependencies and sequences. By focusing on these crucial features, the system enhances its ability to identify authentic versus fake images, ultimately improving the overall accuracy of manipulation detection across both media types.

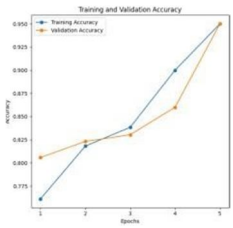
D. Model Creation

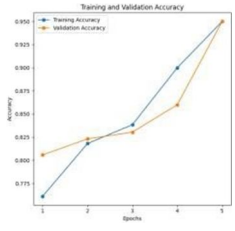
Model creation using the RNN algorithm involves feeding sequential data, like audio or video frames, into the network, which processes the information over time by maintaining memory of previous inputs. The RNN captures temporal dependencies in the data, making it ideal for tasks like deepfake detection. After training on labeled data, the model learns patterns that help classify the input as real or fake based on both past and current information.

V. RESULT AND DISCUSSION

A. Accuracy

Accuracy is a fundamental metric used to evaluate the performance of a classification model, providing insight into its overall effectiveness. It quantifies the proportion of correctly classified instances— comprising both true positives (TP) and true negatives (TN)—in relation to the total number of cases assessed. This means that accuracy reflects the model's ability to correctly identify both the positive and negative classes.





$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP = True Positives (correctly predicted positive cases)
- TN = True Negatives (correctly predicted negative cases)
- FP = False Positives (incorrectly predicted positive cases)
- FN = False Negatives (incorrectly predicted negative cases)

While it is a straightforward measure that can indicate how well a model performs across all classes, it can sometimes be misleading, especially in cases of imbalanced datasets where one class significantly outnumbers the other. In such scenarios, a model could achieve high accuracy simply by favoring the majority class, even if it fails to identify the minority class effectively.

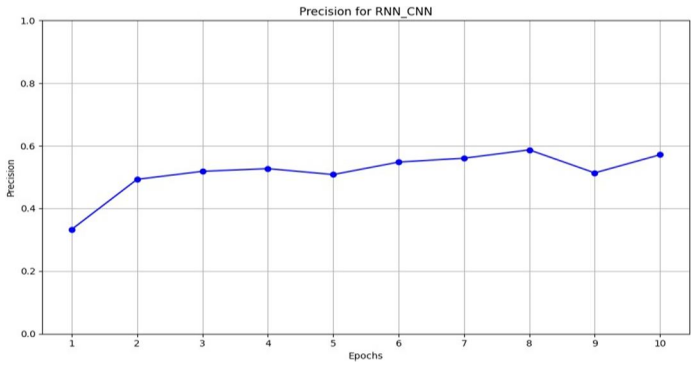
B. Precision

Precision, often referred to as Positive Predictive Value, quantifies the accuracy of the positive predictions made by a classification model. It is calculated by taking the number of true positive results (the cases correctly identified as positive) and dividing it by the total number of predicted positive cases (the sum of true positives and false positives). This metric is crucial in scenarios where the cost of false positives is high, as it helps assess the reliability of the model's positive predictions.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Explanation

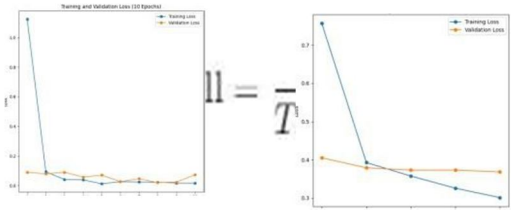
- True Positives (TP): These are the instances where the model correctly predicts a positive class.
- False Positives (FP): These are the instances where the model incorrectly predicts a positive class, meaning the actual class is negative.



A high precision indicates that when the model predicts a positive case, it is likely to be correct, thus instilling confidence in the predictions made by the model. For example, in medical diagnostics, high precision ensures that patients identified as having a disease truly have it, minimizing unnecessary stress and treatment for those who do not.

C. Recall

Recall, also known as Sensitivity or True Positive Rate, is a vital metric in the evaluation of classification models, particularly in fields where identifying positive cases is critical, such as healthcare and fraud detection. It measures the proportion of true positive results in relation to the total number of actual positives. Essentially, recall indicates how well a model can correctly identify positive instances among all instances that belong to the positive class. This makes it particularly important in situations where failing to detect a positive case could lead to serious consequences, such as overlooking a medical condition in a patient.



In this formula, TP (True Positives) refers to the instances where the model accurately predicts a positive case, while FN (False Negatives) represents the actual positive instances that the model fails to identify. A high recall value is indicative of a model's ability to successfully identify a large proportion of the positive cases, which is especially important in applications like disease detection, where it is crucial to minimize the number of missed diagnoses. In contrast, a low recall value signals that the model is failing to capture many of the true positives, which could result in significant oversight and negative outcomes in critical areas.
However, it is essential to understand that there is often a trade-off between recall and precision. While high recall is desired in contexts where identifying as many positive cases as possible is crucial, it can sometimes lead to a higher rate of false positives, thereby affecting precision. Consequently, models need to be carefully calibrated based on the specific needs of the application.
In medical diagnostics, for instance, achieving a high recall is often prioritized to ensure that most patients with a condition are correctly identified, even if it means accepting some degree of false positives. Balancing recall with other metrics like precision and accuracy is crucial for developing robust and effective classification models.

D. F1 Score

The F1 Score is a critical metric in evaluating classification models, especially in scenarios where class distributions are imbalanced. By being the harmonic mean of precision and recall, it offers a single score that encapsulates both the model's accuracy in predicting positive cases (precision) and its effectiveness in identifying all actual positive cases (recall). This balance is particularly valuable in contexts such as medical diagnostics or fraud detection, where failing

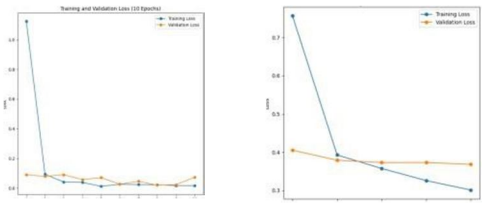
$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

To identify true positives can have significant repercussions. The F1 Score helps to mitigate the impact of any bias toward the majority class, ensuring that the model's performance is not solely assessed on accuracy but also on its ability to capture the nuances of the data.

This formula illustrates how the F1 Score incorporates both precision and recall into its calculation. When dealing with imbalanced datasets, where one class is significantly more prevalent than the other, relying solely on accuracy can be misleading. For instance, a model that predicts all instances as the majority class may achieve high accuracy while failing to identify any instances of the minority class. The F1 Score, by contrast, ensures that both precision and recall are considered, providing a more comprehensive measure of a model's performance, particularly in applications where false negatives are costly or dangerous. Thus, it serves as a critical indicator for models that require a nuanced understanding of class predictions.

E. Loss

The loss function is a crucial component in deep learning models, as it quantifies the discrepancy between the predicted outputs and the



Actual target values, guiding the model's optimization process. In our proposed system, which integrates VGG19 for fake image detection and a CNN-RNN hybrid model for fake audio detection, distinct loss functions are utilized for each component to ensure precise evaluation and learning. For the image analysis using VGG19, we employ Categorical Cross-Entropy Loss, which is well-suited for classification tasks involving discrete classes, such as distinguishing between real and fake images. This loss function penalizes incorrect predictions by comparing the predicted probability distribution of the model to the true class labels, thereby encouraging accurate predictions.

$$\text{Loss}_{\text{image}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Where:

- y_i is the true label (1 for fake, 0 for real).
- $\hat{y}_i$ is the predicted probability of the image being fake.
- N is the number of classes (in this case, 2).

$$\text{Loss}_{\text{audio}} = -\frac{1}{M}\sum_{i=1}^{M} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Where:

- M is the number of audio samples.
- y_i is the true label for each sample.
- $\hat{y}_i$ is the predicted probability of the audio being fake.

On the other hand, for the audio analysis handled by the CNN-RNN hybrid model, we use Binary Cross-Entropy Loss, appropriate for binary classification tasks like detecting fake versus real audio. This loss function evaluates the prediction accuracy by measuring the difference between the predicted probabilities of the audio being fake and the actual labels, penalizing deviations from the true labels. By focusing on both spatial and temporal features in the audio data, this loss ensures comprehensive learning of the audio patterns. To achieve an overall balanced optimization, the system combines these two loss functions, often by calculating a weighted sum of the image and audio losses. This approach allows the model to learn effectively from both modalities, ensuring robust detection of deep fakes across visual and auditory media.

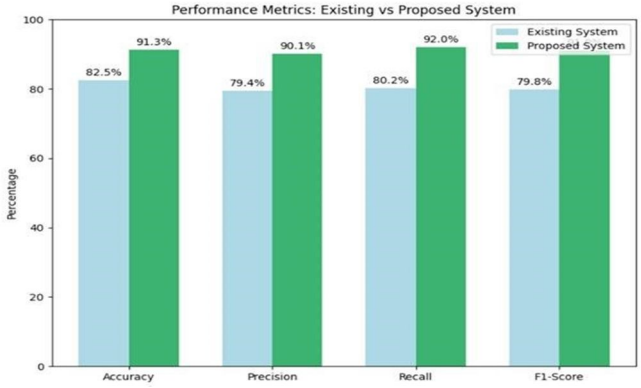
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

F. Prediction

The proposed system uses VGG19 for fake image detection by leveraging its deep convolutional layers and pre-trained weights, which enable it to effectively identify visual artifacts and manipulations in images. VGG19, with its ability to capture hierarchical features, ensures accurate recognition of fake images. For fake audio detection, the system employs a hybrid CNN-RNN model. The CNN component extracts spatial features from audio spectrograms, helping to identify patterns indicative of manipulation. The RNN component captures temporal dependencies in audio signals, ensuring a robust analysis of audio authenticity. This combination allows for accurate and efficient detection of both visual and auditory deep fakes. The VGG19 model specializes in processing image data, while the CNN-RNN hybrid excels in analyzing the sequential nature of audio. By integrating these models, the system provides a comprehensive solution for detecting manipulations in both image and audio content. This dual approach enhances the overall reliability and effectiveness of deep fake detection.

G. Comparison Graph

The proposed system significantly outperforms the existing approach, achieving 91.3% accuracy compared to the previous 82.5%. This improvement highlights the system's enhanced ability to correctly classify real and fake media. Precision, which measures the system's ability toavoid false positives, increased from 79.4% to 90.1%. This leap underscores the reduced error rate in identifying authentic content as fake.With a recall boost from 80.2% to 92.0%, the proposed system demonstrates superior sensitivity, effectively capturing a higher proportion of actual deepfakes, reducing the risk of overlooked manipulations.The F1-Score, a balanced measure of precision and recall, also rose from 79.8% to 90.1%, reflecting the system’s overall improvement in detecting deepfakes with greater consistency. The substantial gains across all metrics confirm the effectiveness of the proposed deepfake detection framework, reinforcing its potential for robust real-world deployment in combating deepfake threats.



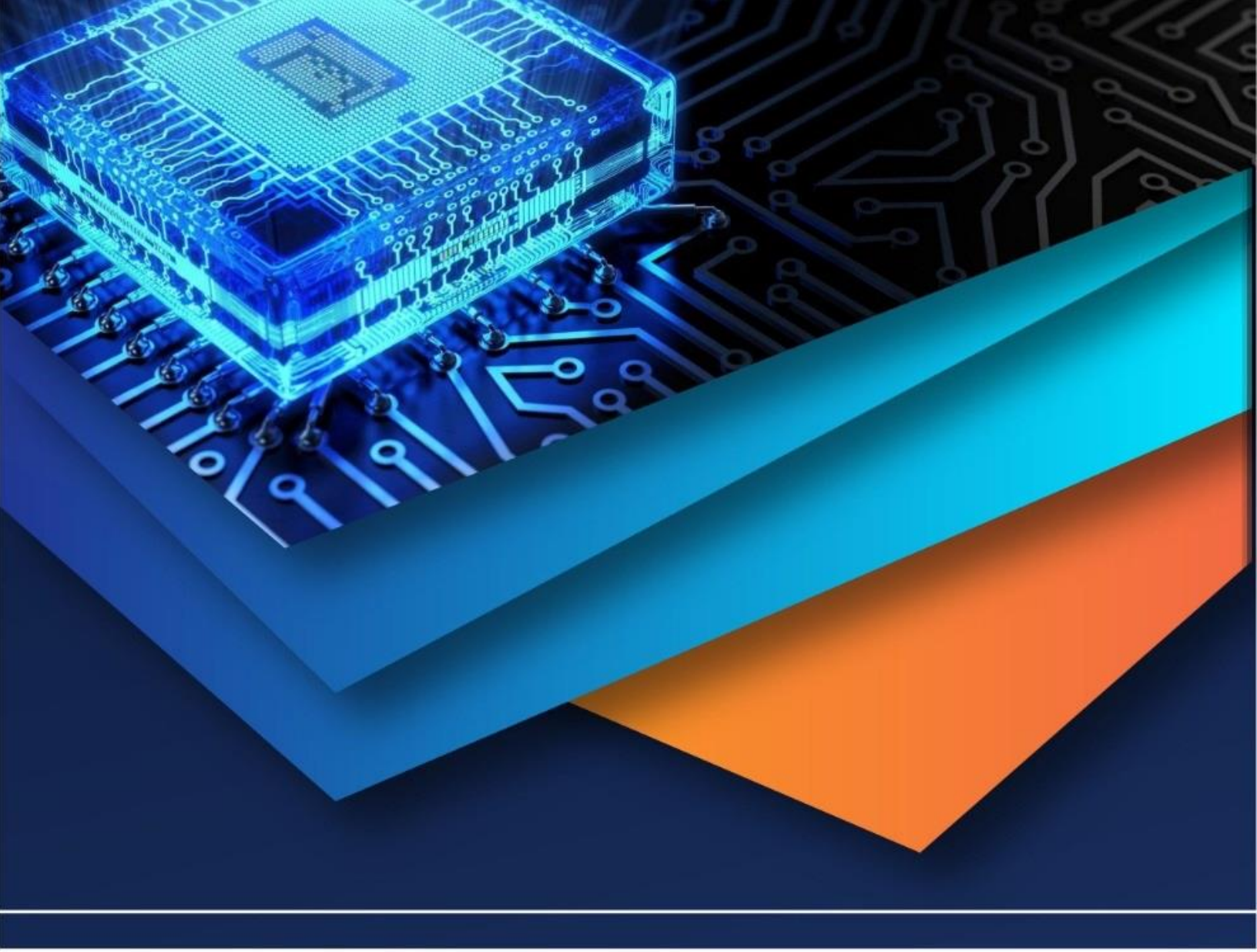
Metric	Existing System	Proposed System
Accuracy	82.5%	91.3%
Precision	79.4%	90.1%
Recall	80.2%	92.0%
F1 Score	79.8%	90.1%

VI. CONCLUSION

The proposed system leverages a hybrid approach with Recurrent Neural Networks (RNN) to detect manipulations in both audio and image content, effectively addressing the unique challenges posed by multimedia forensics. By separately training on audio and image datasets, the system enhances detection accuracy and reliability for each media type. The integration of RNNs allows for better modeling of temporal dynamics in audio data and advanced feature extraction in images, making it a powerful tool for identifying fake or manipulated content. Additionally, the user-friendly web application simplifies the process for users, allowing them to upload and analyze content seamlessly. This system represents a significant advancement in multimedia forensics, expanding detection capabilities while providing a foundation for future research in combating digital manipulations. To further improve the system, future iterations could incorporate additional machine learning models, such as Convolutional Neural Networks (CNNs), to complement the RNN in image analysis. Moreover, expanding the dataset by incorporating more diverse types of manipulations could help increase robustness.

REFERENCES

[1] Z. A. Baig et al., "Future challenges for smart cities: Cyber-security and digital forensics," Digit. Investig., vol. 22, pp. 3–13, Sep. 2017.
[2] H. Zimmerman, "The data of you: Regulating private industry's collection of biometric information," U. Kan. L. Rev., vol. 66, p. 637, 2017.
[3] A. K. Jain and A. Kumar, "Biometric recognition: an overview," in
[4] Second generation biometrics: The ethical, legal and social context, Springer, 2012, pp. 49–79.
[5] D. Lillis, B. A. Becker, T. O. Sullivan, and M. Scanlon, "Current Challenges and Future Research Areas for Digital Forensic Investigation INVESTIGATION," no. c, 2016.
[6] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework," in 2006 IEEE Odyssey-The Speaker and Language Recognition Workshop, 2006, pp. 1–8.
[7] A. Saleema and S. M. Thampi, "Voice Biometrics: The Promising
[8] Future of Authentication in the Internet of Things," in Handbook of Research on Cloud and Fog Computing Infrastructures for Data Science, IGI Global, 2018, pp. 360–389.
[9] B. Zawali, R. A. Ikuesan, V. R. Kebande, S. Furnell, and A. A-
[10] Dhaqm, "Realising a Push Button Modality for Video-Based
[11] Forensics," Infrastructures, vol. 6, no. 4, p. 54, 2021.
[12] Chunlei Peng, Huiqing Guo, Decheng Liu, Nannan Wang, Ruimin Hu, Xinbo Gao. (2023) Deep Fidelity: Perceptual Forgery Fidelity
[13] Assessment for Deepfake Detection arXiv:2312.04961v1
[14] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, Wei Xia. (2021) Learning Self-Consistency for Deepfake Detection IEEE/CVF International Conference on Computer Vision (ICCV)
[15] Bojia Zi ,Minghao Chang ,Jingjing Chen, Xingjun Ma, Yu-Gang Jiang. (2021) Wild Deepfake: A Challenging Real- World Dataset for Deepfake Detection arXiv:2101.01456v1
[16] Kaede Shiohara Toshihiko Yamasaki. (2022) Detecting Deep fakes with Self-Blended Images IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
[17] Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. (2022) Explaining Deepfake Detection by Analysing Image Matching. arXiv:2207.09679v1
[18] Anubhav Jain, Nasir Memon, Julian Togelius. (2022) A Dataless FaceSwap Detection Approach Using Synthetic Images. IEEE International Joint Conference on Biometrics, IJCB. Institute of Electrical and Electronics Engineers Inc.
[19] Fatima Maher Salman and Samy S. Abu-Naser. (2022) Classification of Real and Fake Human Faces Using Deep Learning.
[20] International Journal of Academic Engineering Research (IJAER) 6 (3).
[21] Tiewen Chen , Shanmin Yang, Shu Hu, Zhenghan Fang, Ying Fu, Xi Wu, Xin Wang. (2024) Masked Conditional Diffusion Model for Enhancing Deepfake Detection. ArXiv, abs/2402.0054.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY