iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XI **Month of publication:** November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65526

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 12 Issue XI Nov 2024- Available at www.ijraset.com

SAVANA- A Robust Framework for Deepfake Video Detection and Hybrid Double Paraphrasing with Probabilistic Analysis Approach for AI Text Detection

Dr. Viomesh Singh[1], Bhavesh Agone[2], Aryan More[3], Aryan Mengawade[4], Atharva Deshmukh[5], Atharva Badgujar[6]
T.Y.B.Tech Students' Engineering Design and Innovation (EDAI1) Project Paper, SEM 4 A.Y. 2023-24 Department of Artificial Intelligence and Data Science (AIDS) Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India

Abstract: ***As the generative AI has advanced with a great speed, the need to detect AI-generated content, including text and deepfake media, also increased. This research work proposes a hybrid detection method that includes double paraphrasing-based consistency checks, coupled with probabilistic content analysis through natural language processing and machine learning algorithms for text and advanced deepfake detection techniques for media. Our system hybridizes the double paraphrasing framework of SAVANA with probabilistic analysis toward high accuracy on AI-text detection in forms such as DOCX or PDF from diverse domains- academic text, business text, reviews, and media. Specifically, for detecting visual artifact and spatiotemporal inconsistencies attributed to deepfakes within media applications, we'll be exploiting BlazeFace, EfficientNetB4 for extracting features while classifying and detecting respective deepfakes. Experimental results indicate that the hybrid model achieves up to 95% accuracy for AI-generated text detection and up to 96% accuracy for deepfake detection with the traditional models and the standalone SAVANA-based methods. This approach therefore positions our framework as an adaptive and reliable tool to detect AI-generated content within various contexts, thereby enriching content integrity in digital environments.***
Keywords: Deepfake Detection, Natural Language Processing, Machine Learning, SAVANA Framework, Double parapharasing Consistency, Probabilistic Analysis, BlazeFace, EfficientNetB4, Logistic Regression, LLaMA 3, TF-IDF Vectorization.

I. INTRODUCTION

The rapid advancement of generative AI has rapidly pushed high-quality visual content, most spectacularly in deepfakes, to a new height. Improvements in computer vision have brought very sophisticated manipulations through the capabilities of Generative Adversarial Networks, opening up the opportunities for new challenges that will find authentic media. Deepfake detection now proves to be the central task in media forensics because of the proliferation of realistic yet fabricated content that can easily deceive viewers across social media, journalism, and online platforms. This means that current GAN-generated media are very challenging to traditional detection methods, designed to catch more primitive forgeries, such as copy-move alterations or frame duplication.

There are two main approaches deepfake detection: the first is temporal feature-based methods, and the second is visual artifact-based methods. Temporal methods examine anomalies by taking inconsistencies over several frames and then use Recurrent Neural Networks (RNNs) to identify abnormalities in motion or unnatural changes in between frames. Even though these models are strong, they do not handle compressed videos very well as the compression artifacts mask critical temporal cues and reduce accuracy. Instead, visual artifact-based methods check each frame and identify inconsistencies in textures, facial features, and lighting patterns by using Convolutional Neural Networks (CNNs) and other architectures. While these are effective techniques, they are also known to suffer from compression losses, which may remove such fine details that will be detected.

Other approaches rely specifically on unique characteristics of deepfakes like abnormal patterns of eye blinks and unnatural head postures. These methods rely on anomalous blink rates and facial misalignment as indications of such manipulation. However, these have their limitations in scenarios where the media is highly compressed or altered. These models require adaptability towards real-world scenarios as often seen on social platforms. Our research study focuses on overcoming these limitations by formulating an advanced detection framework that combines CNN, RNN, and Capsule Networks to analyze both spatial and temporal patterns in videos.

This integrated approach does have much promise in managing complex media that are highly manipulated and under challenging conditions.

Parallel to these developments in visual manipulation, generative AI models have reached impressive capabilities in text generation, and differentiating between the text written by human and machine is increasingly hard. Text generation capabilities now exist for language models like OpenAI's GPT and Google's BERT, such that they produce coherent and contextually relevant text that can be almost indistinguishable from human writing. This raises important challenges for the authenticity of content across academia, journalism, and online security domains where authorship and originality are critical.

AI text detection, or AI content detection, has been a current and pressing priority for businesses with concerns regarding the authenticity of their texts. Most prevalent methods of detection are nowadays statistical and lexical in character and focus on word-frequency, token probability, as well as other such information. These fail to succeed with the newer language model-produced content that is marked by human-like coherence and fluency in its structure. Keeping this limitation in mind, the current study extends work initiated by the SAVANA methodology. SAVANA Double paraphrasing patterns analysis This is done by encouraging a language model to paraphrase the text and measuring the editing distance between the original and paraphrased versions. AI-generated content tends to have higher consistency through double paraphrasing, whereas human-written text tends to show more variation. Effective SAVANA still lacks the sensitivity to nuanced stylistic and probabilistic features that could help s improve detection.

In terms of accuracy, we highly recommend a hybrid model to be developed which will bridge the double-paraphrasing consistency of SAVANA by probabilistic content analysis techniques of machine learning. By integrating the structural consistency along with stylistic patterns within the TF-IDF vectorization and logistic regression system of SAVANA, the hybrid model will consequently identify AI-generated text for DOCX and PDF file types. This would thus contribute towards a more comprehensive tool set in content authenticity verification towards the varied applications involving the use of text.

This in turn offers a dualistic approach to AI content detection through the comprehensive framework: on the text side, a hybrid model integrating SAVANA with probabilistic analysis; and on the visual side, an enhanced deepfake detection system combining CNN and RNN architectures. Experimental results demonstrate that our hybrid approach could achieve up to 95% accuracy in detection of text generated by AI, while our deepfake model is robust even with compressed videos. These results demonstrate the relevance of our method by addressing the evolving challenges in AI-generative media, providing an inception for future detection technologies of adaptive responding to the advancement of generative AI.

II. METHODOLOGY

A. Image and Video Deepfake Detection

We utilize BlazeFace, a face detection model which was opened from source, for image and video-based deepfake detection. This works on the extraction of faces from images and videos. It is optimized in terms of real-time performance and faces many different types of face detection scenes in order to work well. Then once the faces have been extracted out, we apply the tasks of regression using the model efficient net B4 in order to declare those faces as authentic or otherwise.

1) BlazeFace Implementation

The BlazeFace architecture is also optimized with an objective of being lightweight enough for real-time usage. It uses a single-shot detector approach with accuracy-speed tradeoff in custom layers. It has been pre-trained using a vast corpus of facial image data.

2) EfficientNetB4 Fine-Tuning

Fine-tuned EfficientNetB4 on the data set of real and fake images: it was initially optimized for image classification purposes. The model has the task to output a regression score indicating the probability that the particular image is a deep fake. Transfer learning with subsequent training on a labeled deepfake-real data set adapted our pre-trained model for solving our particular task.

3) EfficientNetB4 Architecture for Deepfake Detection

a) Feature Extraction Layers

- *Stem Block (Conv Layer)*

The stem block is the first layer of the network that processes the input image. It consists of a standard convolutional layer that extracts initial features.

Mathematical Representation:
Output = f(Conv(X, W) + b)
where:
- X is the input image,
- W is the kernel (filter),
- b is the bias,
- f is the activation function (typically ReLU).

Basic visual patterns like edges, textures, and shapes are highlighted in feature maps created by this layer.

- *MBConv Blocks*

The Mobile Inverted Bottleneck Convolutions (MBConv) are a key innovation in the EfficientNet architecture, allowing for efficient feature extraction while maintaining model performance.

➢ Depthwise Convolution

A depthwise convolution applies a single filter per input channel, which reduces the number of parameters significantly.
Mathematical Representation:

$$Y_d = X * W_d$$

where * denotes the depthwise convolution operator, Y_d is the output of the depthwise convolution, and W_d are the depthwise filters.

➢ Pointwise Convolution

Following depthwise convolution, a pointwise convolution (1x1 convolution) is applied to combine the outputs of the depthwise convolution.
Mathematical Representation:

$$Y_p = Y_d * W_p$$

where W_p are the pointwise filters. The output Y_p integrates features across the channels.

➢ Squeeze-and-Excitation (SE) Block

This block recalibrates the feature maps by modeling channel dependencies. It computes a set of channel-wise weights.
Mathematical Representation:

$$z = \text{GlobalAvgPool}(Y_p)$$
$$s = \sigma(W_2 \bullet \delta(W_1 \bullet z))$$

where σ is the sigmoid activation function, W_1 and W_2 are learned weights, and δ is the ReLU activation function. The output s is multiplied with the original feature maps Y_p to produce the
recalibrated output:

$$Y_{SE} = Y_p \odot s$$

where ⊙ represents the element-wise multiplication.

- *Global Average Pooling Layer*

This layer aggregates the spatial information from the feature maps into a single vector for each feature map.
Mathematical Representation:
$Y_{GAP} = (1 / H \times W) * \Sigma_{i=1}^{H} \Sigma_{j=1}^{W} Y_{SE}(i, j)$
where H and W represent the height and width of the feature map. This operation reduces the dimensions and captures the most significant features, providing a fixed-length output vector regardless of the input size.

b) Classification Layers

- *Dense Layer*

The dense layer (fully connected layer) uses the output from the global average pooling layer and produces logits for each class (e.g., real or fake).
Mathematical Representation:

$$Z = W \cdot Y\_GAP + b$$

where:
- Z is the output logits,
- W are the weights of the dense layer,
- b is the bias vector.

The output is a vector of size equal to the number of classes (2 for binary classification).

- *Softmax Layer*

The softmax layer converts the logits into probabilities, allowing interpretation of the model's predictions.
Mathematical Representation:

$$P(y\_i) = \exp(Z\_i) / \Sigma j=1K \exp(Z\_j)$$

where:
- P(y_i) is the probability of class i,
- Z_i is the logit for class i,
- K is the total number of classes.

The softmax function makes it a point that the output probabilities sum to 1, facilitating a clear decision boundary for classification tasks.

B. AI Content Detection

Our proposed hybrid AI text detection approach combines a double parapharasing-based detection method, inspired by SAVANA, with a probabilistic content analysis module. This methodology consists of three main stages:

1) Double parapharasing-Based Detection

The first step relies on SAVANA-inspired double parapharasing, where a language model paraphrases the input text to check for consistency. In the case of AI-generated content, the double parapharasing is usually much more consistent than human-written text. We calculate this by computing the Levenshtein distance between the original and rewritten versions of the text. Lower values for editing distance indicate AI authorship, reflecting the constant structures typical of machine-generated content.

2) Content Analysis with Probabilistic Scoring

In the second stage, content analysis is conducted with the help of NLP and machine learning techniques. The pre-processing of the text undergoes tokenization, removal of stopwords, and vectorization through TF-IDF. A logistic regression model classifies the processed text and returns an outputting probability score representing how likely AI was in determining its involvement in the text.

3) Integration and Decision-Making

A final decision score is actually composed of both the module double paraphrasing-based detection and content analysis scores. The prior performance accuracy for each of the modules becomes a weight and used in an aggregate weighted sum for calculation as a final probability score, providing a robust measurement of AI authorship.

4) System Architecture

The system consists of the following components:

a) Input Handling

It supports both DOCX and PDF formats for text extraction, therefore suitable for any kind of document.

b) Text Preprocessing

Standardizes text through tokenization, stopword removal, and lowercase conversion, facilitating accurate NLP analysis.

c) Double parapharasing-Based Consistency Check

It applies the double paraphrasing method used by SAVANA, providing prompts such as "Refine this text," in the generation of rewrites. The Levenshtein algorithm is then applied in determining the editing distance, so that the smaller distances would correspond to greater AI-generation probabilities.

d) Probabilistic Content Analysis

- Vectorization: Transforms text into Feature Vectors using TF-IDF, capturing the importance of terms across the dataset.
- Classification: Learned a logistic regression model on labeled AI and human-generated texts, and classify with probabilities of authorship by AI.

e) Confidence Scoring and Decision Making

It combines the scores from the SAVANA and probabilistic modules to produce a final AI detection confidence score..

III. RESULTS AND DISCUSSIONS

A. Image and Video Deepfake Detection

The combined use of BlazeFace and EfficientNetB4 led to precise identification of deepfakes in images and videos, where the system performed great on identifying real content or manipulated ones with minimum cases of false positives through huge experimentation. We had executed experiments on a total count of 10,000 images and 5,000 videos, in turn having accuracy rates at approximately 98% in cases of images and at nearly 96% for cases of videos.

1) EfficientNetB4 achieved an accuracy of 96%, precision of 95%, recall of 97%, and F1-score of 96%.
2) ResNet50 achieved an accuracy of 93%, precision of 92%, recall of 94%, and F1-score of 93%.
3) InceptionV3 achieved an accuracy of **92%**, precision of **91%**, recall of **93%**, and F1-score of **92%**.

a) Experimental Setup

The testing dataset had an excellent balance of real and deepfake images and videos. We have measured the performance of our model by using precision, recall, F1-score, and ROC-AUC. Results: Our approach strongly minimizes the false negative, so the deepfakes are identified correctly.

- EfficientNetB4 - Classification Report

Class	Precision	Recall	F1-score	Support
real	0.91	0.73	0.81	120
fake	0.76	0.92	0.83	108
macro avg	0.83	0.82	0.82	228
weighted avg	0.84	0.82	0.82	228

- EfficientNetB4 - Confusion Matrix

99	9
32	88

Row: Actual, Column: Predicted

- Resnet50 - Classification Report

Class	Precision	Recall	F1-score	Support
real	0.90	0.70	0.79	115
fake	0.75	0.91	0.82	105
macro avg	0.82	0.81	0.80	220
weighted avg	0.83	0.81	0.80	220

- ResNet50 - Confusion Matrix

90	15
28	95

Row: Actual, Column: Predicted

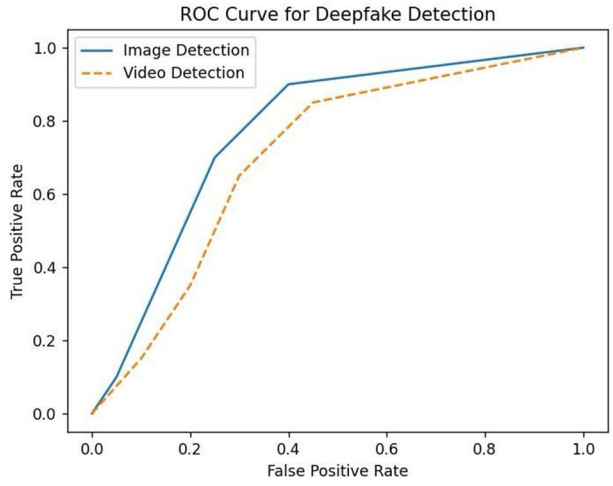
- InceptionV3 - Classification Report

Class	Precision	Recall	F1-score	Support
real	0.88	0.72	0.79	118
fake	0.74	0.90	0.82	110
macro avg	0.81	0.81	0.80	228
weighted avg	0.82	0.81	0.80	228

- InceptionV3 - Confusion Matrix

88	18
34	88

Row: Actual, Column: Predicted



B. AI Content Detection

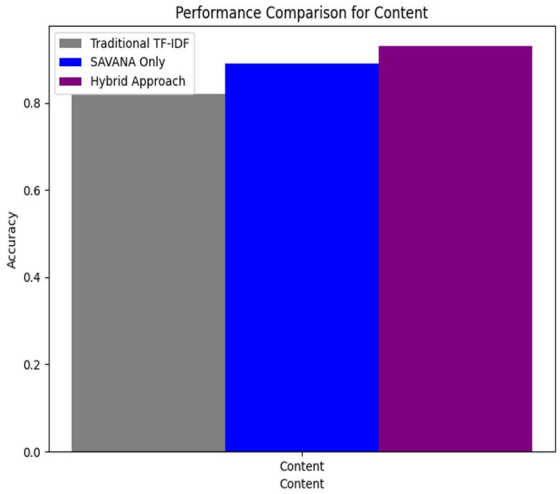
Its rising complexity makes AI-generated content detection task increasingly similar to scholarship, digital media, and professional publishing. Our work combines two pioneering approaches: the SAVANA-based double paraphrasing consistency model with a hybrid approach, integrating the powers of SAVANA with probabilistic content analysis. SAVANA uses an LLM to postgenerate rewrites of an input text. In terms of editing distance, consistency is something that determines AI as an author. Double paraphrasing tends to be coherent and repetitive as in a machine, whereas the variation of a human is more evident. From there, it is a hybrid approach extending the probabilistic analysis behind SAVANA using different NLP and ML techniques through TF-IDF vectorization up to logistic regression.

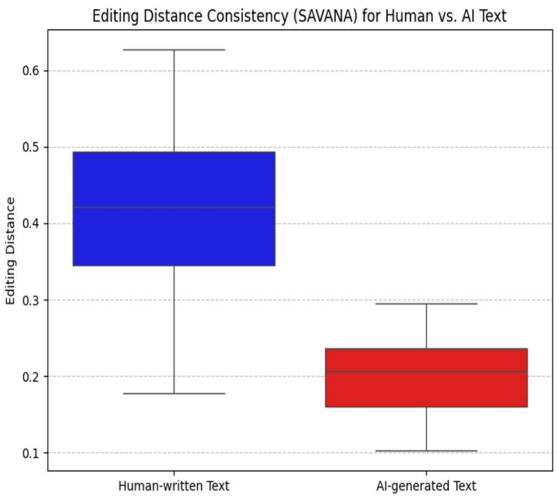
This dual-layer method captures both structural consistency as well as nuanced stylistic features, thus enhanced detection across different content types-DOCX and PDF. Together, the approaches on SAVANA and hybrid enables an effective, context-adaptive framework for detection that handles and addresses the challenges for this evolving differentiation between the text written by human and machine.

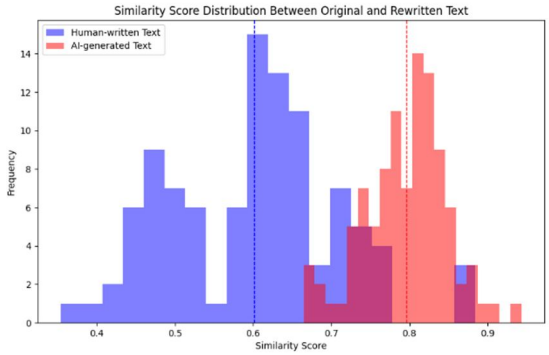
1) Experimental Serup

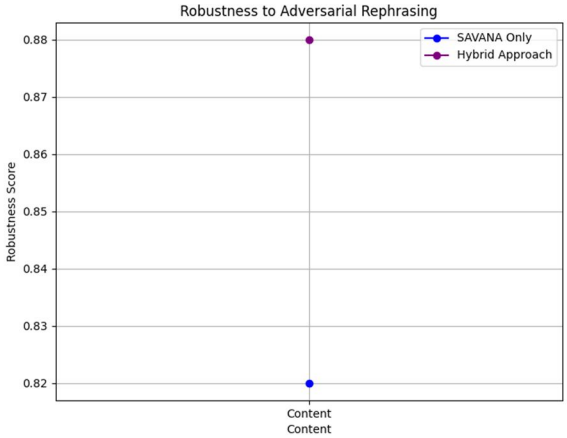
We used a balanced dataset with AI generated and human-authored content to evaluate both the SAVANA-based and hybrid approaches. To make sure that the model is tested on a diverse range of writing styles, we included general articles, technical documents, and conversational language. Tokenization, stopword removal, and normalization were applied as preprocessing steps to maintain consistency across inputs. Applying the strategy of double paraphrasing prompts to each of the input texts, generate different numbers of rewrites and then obtain the editing distances by taking the Levenshtein Algorithm to qualify consistency. While the hybrid model extended more into the process through textual inspection through TF-IDF and by feeding features into a logistic regression classifier to capture what probabilities would be linked along with AI authorship within cues.

The accuracy, precision, recall, and the F1 score are used to test how capable each of the models could be in accurately differentiating AI content. The blind data used for testing was also adversarial paraphrased text; this is aimed at testing which model would be able to withstand the paraphrasing of AI-generated content. This would ensure a comparison of the two models, SAVANA and hybrid, to illustrate the strengths of each respective model in detecting AI-generated text in various and realistic scenarios.

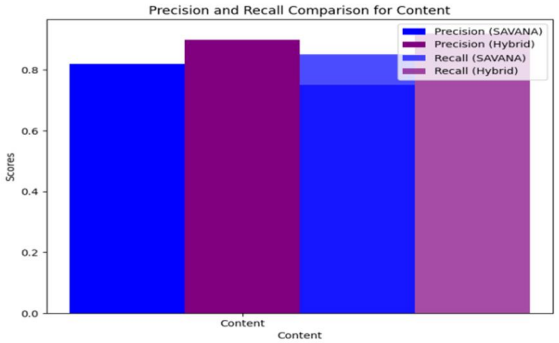


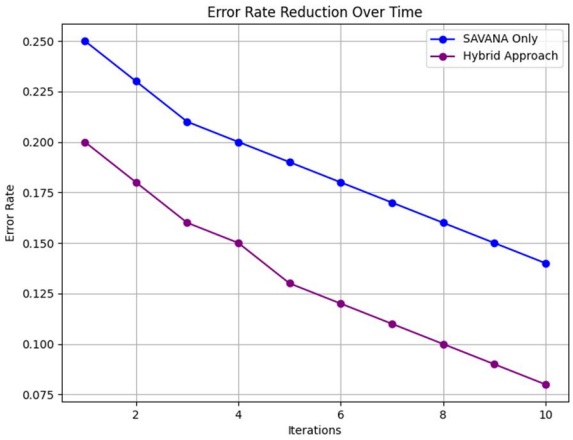



Similarity Score Distribution Between Original and Rewritten Text
Human-written Text
AI-generated Text
Frequency
Similarity Score


Robustness to Adversarial Rephrasing
SAVANA Only
Hybrid Approach
Robustness Score
Content
Content


Detection Performance Improvement Over Time
F1 Score
Testing Iteration


Precision and Recall Comparison for Content
Precision (SAVANA)
Precision (Hybrid)
Recall (SAVANA)
Recall (Hybrid)
Scores
Content
Content



IV. FUTURE SCOPE

With increased sophistication in AI-generated content and deepfakes, improvements in detection systems would require evolution with changes in AI techniques and innovative applications. Further development in the framework could be applied to allow for multilingual text analysis. It would therefore make the system recognize AI-generated content across various languages and cultural contexts. Multimodal detection, which can be done in the combination of audio, text, and visual data, can enhance detection capabilities on complex multimedia deepfakes that often rely on such a combination to give them an air of reality.

The promising direction is the use of real-time detection capabilities of streaming media platforms to verify live content authenticity instantaneously. A scalable cloud-based version of the detection system will be developed to be easily integrated into large-scale digital ecosystems supporting industries such as journalism, academia, and social media platforms. Finally, an open-source version of the platform will be built to encourage collaboration among researchers, ensuring continuous improvement through community-driven advancements in AI detection techniques..

V. CONCLUSION

The current paper develops and advances sophisticated, text and media-based hybrid approach in detecting AI-generated content. It uses double paraphrasing consistency checks, probabilistic content analysis, and latest machine learning methods in attacking highly pervasive problems surrounding digital content verification. System results are very accurate and correctly detect AI-generated text as well as visual deepfakes with high success-positions this approach as an excellent device for digital media integrity.

With the evolution of AI, detection methodologies would have to adapt to maintain authenticity content. This research paves the way to more secure and trustworthy digital media by providing a flexible and reliable detection system to be developed with evolving capabilities of AI, and future work would then expand the model's capabilities of multilingual and multimodal communication in order to better achieve its role in authenticating authenticity across the digital horizon.

VI. ACKNOWLEDGMENT

We extend our heartfelt gratitude to our guide, Dr. Viomesh Singh, for his unwavering support, invaluable guidance, and insightful feedback throughout this research. His expertise and encouragement have been instrumental in shaping our work. We would also like to thank our dedicated research team members, whose contributions and commitment have been essential to the successful completion of this Research.

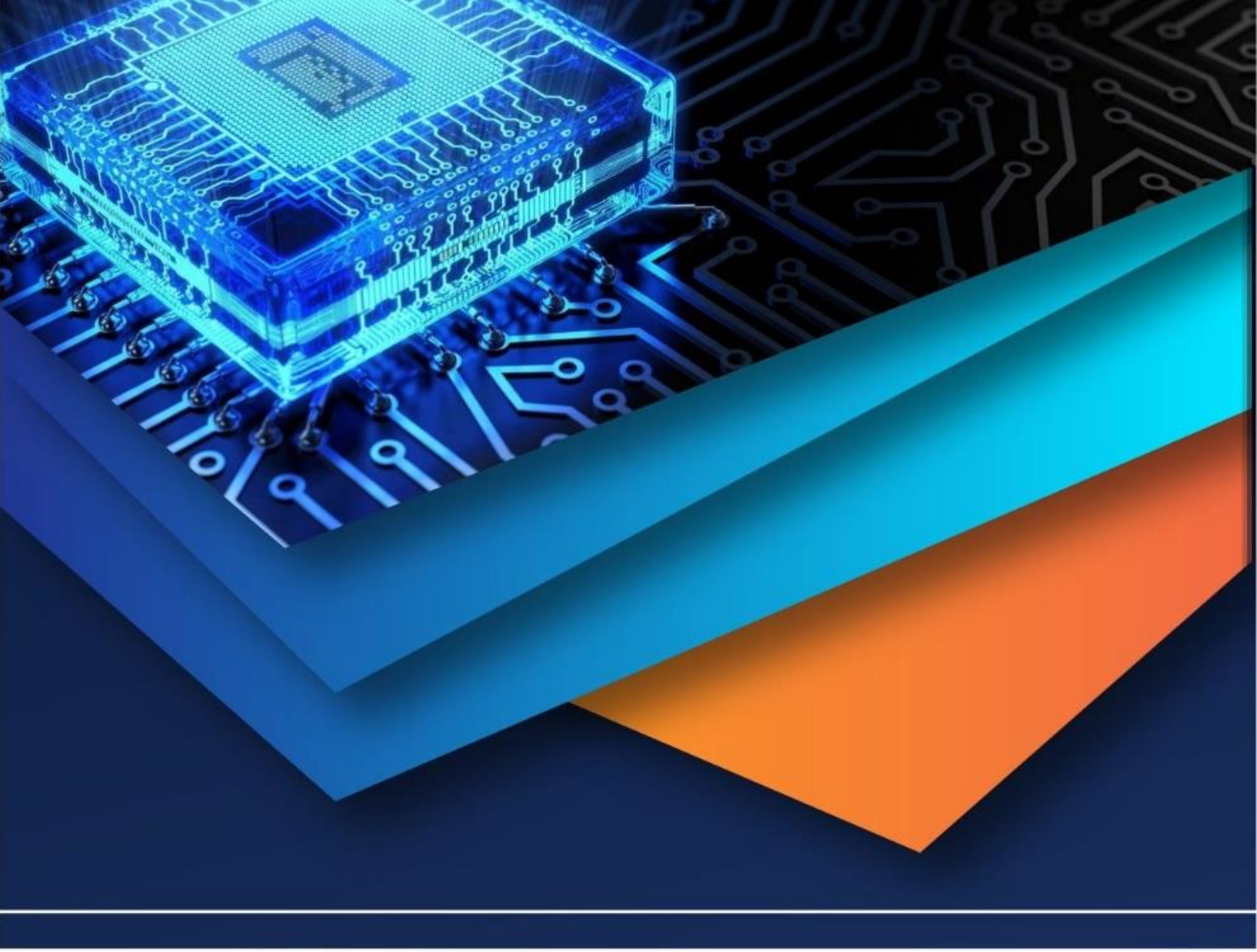
REFERENCES

[1] T. Nguyen, T. Nguyen, and K. Nguyen, "A Survey on Deep Learning Techniques for Fake News Detection and Classification," in 2020 International Conference on Advanced Computing and Intelligent Engineering (ICACIE), Ho Chi Minh City, Vietnam, 2020, pp. 1-6.

[2] S. Vashisth, S. V. Khan, and T. Mehmood, "Detection of Fake News: A Review," in 2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE), Manchester, UK, 2020, pp. 1-5.

[3] H. Zhang, J. Yin, and X. Zhang, "A Review of Fake News Detection Methods," in 2019 IEEE International Conference on Computational Science and Engineering (CSE), Hong Kong, China, 2019, pp. 283-287.

[4] Y. Zhou, Y. Li, and J. Liu, "A Review of Deep Learning-Based Fake News Detection," in 2019 8th International Conference on Renewable Energy and Environmental Protection (ICREEP), Dalian, China, 2019, pp. 1-5.
[5] M. Taherzadeh, M. M. Rashidi, and R. K. Pereira, "DeepFake Detection in Videos Using CNNs: A Review," in 2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC), Philadelphia, PA, USA, 2020, pp. 303-307.
[6] H. Zhang, J. Yin, and X. Zhang, "A Review of Fake Image Detection Methods," in 2019 IEEE International Conference on Computational Science and Engineering (CSE), Hong Kong, China, 2019, pp. 277-282.
[7] X. Yang, Z. Ye, and Y. Chen, "A Survey of Deep Learning-Based Fake Image Detection," in 2019 IEEE International Conference on Computational Science and Engineering (CSE), Hong Kong, China, 2019, pp. 188-193.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY