International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12– No.42, November 2023 – www.ijais.org

Enhancing the Fight against Social Media Misinformation: An Ensemble Deep Learning Framework for Detecting Deepfakes

Ejike Joseph Aloke
M.Sc. Student,
Department of Computer Science,
Faculty of Computing, Federal University Lafia
P.M.B 146, Lafia, Nasarawa State, Nigeria.

Joshua Abah
Senior Lecturer,
Department of Computer Science,
Faculty of Computing, Federal University Lafia
P.M.B 146, Lafia, Nasarawa State, Nigeria.

ABSTRACT
Deepfakes are synthetic media that replace someone's action (source person) with another (target person). Images deepfakes, commonly known as "visual deepfakes," depict a complicated and contentious high-tech avant-garde phenomenon in the sphere of digital trickery and artificial intelligence. These are highly deceitful and computer-based distortions of static images, commonly photographs, where the appearance of a single individual is painstakingly superimposed onto another in a highly sophisticated manner that seems to be real. Image Deepfakes are easy to generate due to easy access to open-source deepfake generation software applications such as FakeApp. Once it is generated, social media becomes its marketplace where it is easily distributed to engage and deceive millions of users. Most research in this area focuses on using a single deep-learning algorithm on a small dataset in the development of the deepfakes detection model. Therefore this research work is focused on building a robust and efficient deepfakes image detection model using a publicly available dataset from Kaggle comprising one hundred and forty thousand (140,000) images. The model was developed using Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Feed Forward Neural Network (FFNN), and Gated Recurrent Unit (GRU). To make the model more robust and efficient, the ensemble technique was employed to ensemble the individual models and an accuracy of 94.91% was achieved.

General Terms
Image Recognition, Security.

Keywords
Deepfakes, Misinformation, Social media, Ensemble Method.

1. INTRODUCTION
The appellation 'deepfake' emanated from two words 'deep learning' and 'fake' [1]. Through the use of deep learning algorithms, software developers have created software capable of manipulating images [2]. These manipulated contents appear to be almost indistinguishable from the original content.

Social media platforms have become part of society's daily routine, as they depend on it for quick information. News agencies, political leaders, and celebrities share personal and public information through their social media handles. Years past, society depended hugely on Radio, TV and newspapers for information [3]. Currently, millions of users now depend on social media for easy and swift information. This information can either be real or fake. Facebook and Twitter have millions of users who create personal and professional accounts for information sharing; due to privacy policies, most of this information does not pass through proper scrutiny [4]. This has given rise to social media misinformation.

Image deepfakes are often underrated but it is still as menacing as the rest [5]. In the first quarter of 2023, a picture of Pope Francis putting on an expensive Balenciaga cloth was all over social media [6]; the image generated a lot of traffic and criticism on Facebook and Twitter simultaneously. Many believed that the image was real but it was deepfaked. Deepfakes images have been in existence for years; as early as 1860, although as of then it was not called deepfake. The formal Vice President of the United States John Calhoun's portrait was painstakingly altered and the head was replaced with the head of the former president of the United States Abraham Lincoln [7]. Then such alteration was done by removing the head of the source person, and replacing it with the target person, repainting and modifying both images to look alike [8]. The above method is known as a manual or handcrafted method.

Presently, computer graphics and the emergence of deep learning algorithms have further made the process easier. Autoencoders and Generative Adversarial Networks (GAN) play a vital role in the creation of deepfake images [9]. Reddit is a popular American social media platform where users are permitted to share multimedia content and interact with one another. In 2017 one of its users with an account named "deepfake" uploaded a succession of computer-perverted fake multimedia content of popular American female celebrities [10]. The faces of the celebrities in the multimedia content were interchanged with faces in pornographic content. Certainly, the prime prey of deepfake generators is public influencers and political leaders. The US Speaker Nancy Pelosi 2019 faced a deepfake attack where she was made to appear chaotic and intoxicated [11]. When the fake content was reposted on Facebook it gained over 2.2 million viewers in less than two days. The above and many other happenings gave deepfakes popularity.

The oldest Software application developed for deepfakes distortion was Video Rewrite [12]. In 2019, an application that was downloadable on Windows and Linux was built solely for the creation of erotic images. This became a peril in society as it created fear and misinformation. People can create erotic images of someone using deepNude without the consent of the

person [13]. FakeApp, FaceSwap, FaceApp, Reface, and ZAO are some of the software developed and used for deepfakes creation. You must not be a computer guru to be able to operate the software as it is so facile to operate, especially with the help of YouTube tutorial videos. The GitHub platform has the deepfaceLab [14] used in generating deepfake content. The major motive behind deepfakes is to misinform and bewilder the receivers. Same 2017, deepfake was used to beget nude images of celebrities; which went viral on social media [10]. Deepfakes have aggravated chaos and blackmailing of notable personalities [15]. Deepfakes can hoodwink the military and security experts through falsified war armament. This study proposes a model for detecting social media image deepfakes which will extenuate misinformation and enhance security. The key contributions of this study are as follows

i. This research work enhances the understanding of the concept of image deepfakes and the models employed in detecting them.

ii. The developed model unveils the framework of an advanced way to develop an image deepfakes detection model.

iii. The research has explored and employed image properties both real and fake from different articles to add more clarity to the detection of image deepfakes.

iv. Multiple deep learning and machine learning algorithms were employed to evaluate the publicly available dataset from the Kaggle online platform. The performance of the image deepfakes detection model was generally measured using a comprehensive confusion matrix that encompasses accuracy, Recall, Precision, and F1 score.

v. This research work implemented transfer learning and ensemble techniques on a large dataset of 70,000 to ensure that the developed model is robust and effective in image deepfake detection.

The rest of the paper is organized as follows: Section (2) focuses on the review of prior research. Section (3) elaborates on the materials and methods employed. Section (4) presents the experimental findings and discusses the study. Section (5) offers conclusions and recommendations for future research.

2. LITERATURE REVIEW

This segment examines numerous research endeavours focused on detecting image deepfakes. Deepfakes are created and spread on social media mainly to deceive the public and cause misinformation. While the term "Deepfake" has gained significant attention, it remains crucial to comprehend the academically accepted definition of this expression.

2.1 Definition

Deepfakes is relatively a new research area, there is yet to be a generally accepted definition. “Synthetic image animation: Application of movement to a static source image through deep neural networks with the goal of creating synthetic visual media of a person” [16]. “manipulated digital media such as images or videos where the image or video of a person is replaced with another person’s likeness” [17]. “deepfakes. . . are created by techniques that can superimpose face images of a target person to a video of a source person to make a video of the target person doing or saying things the source person does . . . deepfakes are artificial intelligence-synthesized content that can also fall into two other categories, i.e., lip-sync and puppet-master” [10]. The next subsection discusses the issues of deepfakes.

2.2 Issue of Deepfakes

Image deepfakes are generated using deep learning algorithms. These images are super sophisticated and barely indistinguishable from the original content. It is capable of causing chaos and misinformation. Deepfakes images can be employed to generate a piece of convincing false news and spread misinformation on social media [18]. Deepfake technology can be employed to generate erotic content by superimposing an individual’s face onto an explicit or erotic image without their consent. Such content can also be used to blackmail or exploit high-profile persons [19]. Deepfakes images can be employed for identity theft or impersonation of individuals, not excluding financial fraud and impersonation of high-profile persons for malicious reasons [20]. Deepfake images can be employed to influence political gatherings by generating a convincing image of popular political personalities saying or doing things they never did [21]. Deepfakes images can be used to bypass facial recognition systems, potentially compromising the security of various systems and devices [22]. The authenticity of image media content keeps decreasing as image deepfakes advance, as it is becoming more difficult to distinguish between real and fake content [23]. Deepfakes proliferation has become a matter of concern as legal entities have to update their rules, laws and regulations. Image evidence needs a proper investigation to avoid passing a wrong judgment [24]. Historical images and photographs can be manipulated by deepfakes images, thereby causing misinformation, especially in digital libraries [15].

The following subsection discusses machine learning related works.

2.3 Review of Machine-Learning-Related Works

Deepfakes has been a research area since its inception; researchers have achieved notable results. In the year 2018, [25] proposed a model to spot computer-manipulated images. The model was trained using a deep neural network called deep forgery discriminator (DeepFD) and it was able to detect deepfake images generated by GAN at an accuracy of 94.7%. In 2019, a model was proposed by [26] to Classify real faces and forged. The classifiers utilized in the training of the model are ShallowNet, VGG-16 and Xception. In the end, they achieved an accuracy of 62%. In 2020, [27] Developed a model to tackle image deepfakes. The model was built using CNN and SVM. After training, the model achieved 94% and 65% on CNN and SVM respectively. The dataset used during the model training was small, as they made use of 2041 images. In the year 2022, [28] proposed a model to detect deepfake images; the model was built using VGG16 and CNN. The model arrived at achieved an accuracy of 94%, although the dataset they used was small and unbalanced; the real images were 1081 and the fake images were 960.

To overcome the challenges of the existing models, this study employed the methodology described in Section 3 below

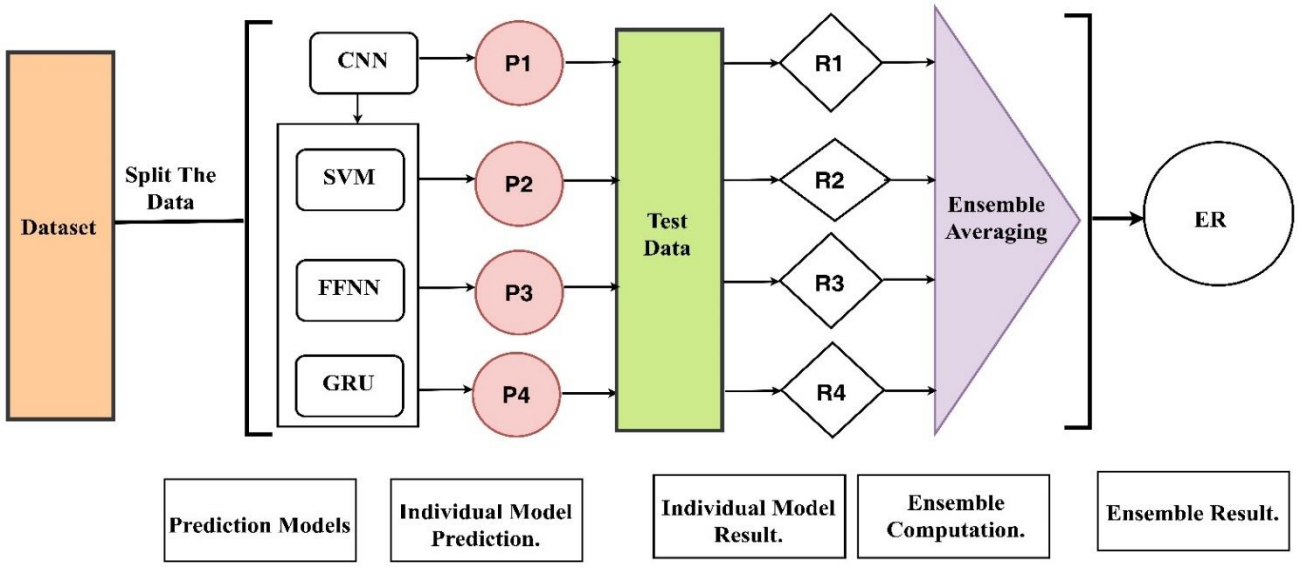

Figure 1: Shows the Ensemble Learning Framework for Deepfakes Detection.

3. METHODS AND MATERIALS

3.1 Proposed Framework

The research introduces a novel approach through the utilization of Ensemble learning, which leverages the strengths of multiple algorithms to mitigate biases that may result from their individual weaknesses. The model framework incorporates a variety of algorithms, including CNN, SVM, FFNN, and GRU. These algorithm selections were based on their relevant characteristics and their prior use in tasks related to image analysis, and deepfakes detection. The primary aim is to develop a robust and high-performing model for the detection of social media image deepfakes. This section will focus on (i) Dataset (ii) Data Preprocessing (iii) Feature Extraction (iv) Model Development (iv) Model Evaluation. Figure 1 shows the ensemble learning framework for the deepfakes detection model. The next subsection focused on data collection.

3.2 Dataset Used

The quality of the dataset used in the development of the model is very important in achieving a good result [17]. The performance of the model relies on the dataset used in the training process [29]. The model development process started by accessing an open-source dataset of images publicly available on Kaggle. It contains a total of 140,000 face images. Half of the faces (70,000) are real from the Flickr dataset collected by Nvidia and the other half (70,000) are fake faces (generated by StyleGAN) that were provided by Bojan [30]. 70,000 images were utilized for the development of the model. The dataset contains three CSV files with features such as id, original_path, label, label_str, and path. Table 1 below describes the features. The next subsection discusses data preprocessing.

Table 1: Dataset Feature Description

S/N	Features	Description
1.	Id	A unique identifier for each image.
2.	original_path	The path to each image in the Kaggle input directory.
3.	label	Is the column that indicates the binary status of each image 0 or 1. 0 for fake and 1 for real.
4.	label_str	This is similar to ‘label’. This column indicates the string value of the image. It shows whether the image is ‘real’ or ‘fake’
5.	Path	This is similar to ‘original_path’. This column indicates the path to the folders that contain the images.

3.1 Data Preprocessing

Data preprocessing is an important phase in the construction of a deep learning model because it ensures that data are cleaned, transformed and arranged in a way that will be suitable for training of the model. At this phase missing values are handled, anomalies are corrected, and data normalization, scaling, augmentation, labelling, splitting, and balancing all took place. The dataset is publicly available and free to use for research purposes. Pandas and numpy libraries in Python programming language were used to manipulate and handle the CSV file

containing the paths and labels of the images. Keras library was used for preprocessing and data augmentation of the images. DataFrame was created for each set Train, validation and test while the Pandas library was used to read the CSV file containing the labels and paths to the images. The images were further resized and normalized to fit into the model. A data visualization tool was employed to check if there was a balance between the fake and real image sub-classes of the train set. The next subsection discusses data splitting.

3.1.1 Data splitting

The dataset was divided into training, validation and testing sets. 50,000 images for training, 25,000 real images and 25,000 fake. 10,000 images, 5,000 real and 5,000 fake were utilized for validation and 10,000 images for testing, 5,000 real and 5,000 fake. The numbers of real and fake images were equal in each subset train, validation and test set. This is to ensure that the model did not suffer data imbalance and to enable the model to learn and converge appropriately. The dataset is suitable for the research because it cuts across various ethnic groups, races, colours, ages, image backgrounds and image accessories such as eyeglasses, sunglasses, hats etc. just like social media images. The split ratio is 71.4%, 14.3%, and 14.3% for training, validation and test sets respectively. Figure 2 shows a few samples of the images. Figure 3 shows the balance between real and fake images. The next subsection focused on the feature extraction.

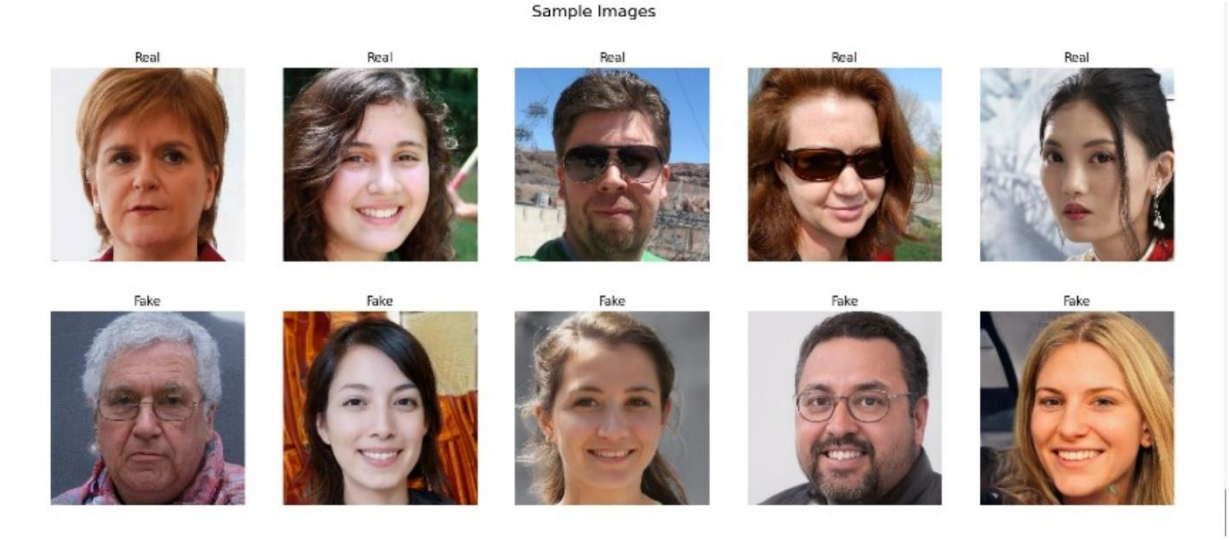

Fig 1: Sample of Image Dataset

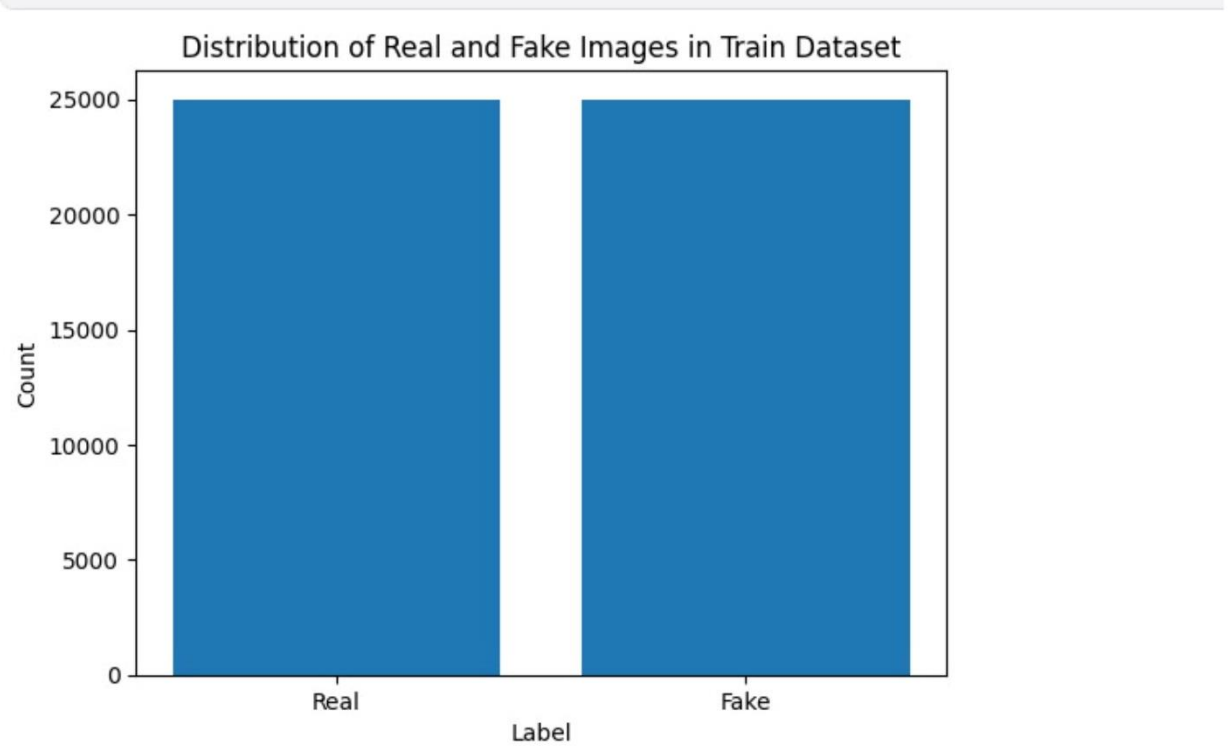

Fig 2: Shows the Balance between Real and Fake Images

3.2 Feature Extraction

In machine learning, feature extraction refers to the process where important information is been extracted from the raw data to generate a more comprehensive informative representation, mostly called features [31]. In the deepfakes image detection model, feature extraction involves the transfiguration of the raw image data into a set of meaningful characteristics and patterns that can be used for the classification process. At this phase, dimensionality was reduced, and patterns were identified, making it compatible and easy for generalization, and interpretation to ensure an efficient model. The quality of images is of importance in this study. The next subsection focused on model development.

3.3 Model Development

Employing multiple classification algorithms in the development of a deepfake detection model can boost accuracy and make the model more vigorous in the detection task [32]. Deepfakes' complexity keeps increasing daily as more sophisticated algorithms emerge and attackers don't relent as they leverage the algorithms to create more misleading deepfakes content to enable them to achieve their ulterior motive. Therefore depending on one classification algorithm will not be strong enough to spot all kinds of image deepfakes [32]. To enable the model to be more proactive and robust in spotting deepfakes image manipulation, multiple algorithms should be used as they can increase the accuracy and general performance of the model [33]. In this phase, Convolutional Neural Networks (CNNs), Support Vector Machines (SVM), Feed Forward Neural Networks (FFNN), and Gated Recurrent Unit (GRU) neural networks were trained and evaluated. After the evaluation of the models, the ensemble technique was utilized to make the model robust.

3.3.1 Classification Algorithms

3.3.1.1 Convolutional Neural Network (CNN)

Is a type of deep learning algorithm that is generally used in the classification and analysis of visual data like images. It has achieved remarkable results in object recognition, image classification, and image analysis [34]. CNNs are suitable for image deepfake detection because they can perform feature extraction on the images used in the classification. The extracted features are learned by the model during the training process; the network simultaneously analyzes the patterns in the image data. CNN further uses the learned expertise to determine which image is real or fake [35]. Moreover, CNNs are known for their ability to work with a large dataset which is very needful in image deepfake detection as voluminous samples are needed for the efficiency and effectiveness of the trained model. Images come in various sizes, colours, and orientations; CNNs have the capacity to handle such diverse images [36]. Social media has millions of users, so it is necessary to utilize an algorithm that can handle diverse and voluminous datasets [37]. The CNN model was trained from scratch, a total of seventy thousand (70,000) images were used, and both the real and fake images were equal in number. The dataset was split into three subsets train set fifty thousand (50,000) images, validation set ten thousand (10,000) and test set ten thousand (10,000) and achieved an accuracy of 94.78%. Table 2 below shows the performance of the CNN Model. The next subsection discusses transfer learning.

Table 2: The CNN Performance Summary

Performance of CNN Model	
Evaluation Metrics	**Performance Results (%)**
Accuracy	94.78
Precision	93.92
Recall	95.76
F1-Score	94.73
Auc Score	98.80

3.3.2 Transfer Learning

Transfer learning allows the transfer of the knowledge gained from a previously learned task to be applied to another similar task [38]. Transfer learning reduces the cost of learning and also helps avoid reinventing the wheel [39]. It is a machine learning method, where a pre-trained model on one task is adapted for a different but related task. Transfer learning is important in image deepfake detection models as it helps in knowledge transfer between models, improves efficiency and generalization, saves time and computational resources, reduces complexity, and enhances performance. After training the CNN model, a transfer learning technique was utilized to train the SVM, FFNN, and GRU models. Below is the discussion of each model:

3.3.2.1 Support Vector Machines (SVMs)

A well-known machine learning algorithm utilized in classification and regression tasks. They are specifically employed in resolving complicated problems with an explicit separation among the various classes. SVMs are efficient in deepfake image detection because of their effectiveness in the separation of various categories of images using hyperplane in a multidimensional space [40]. SVMs are generally employed in image classification such as deepfakes detection, and it has shown a reasonable result when implemented in conjunction with techniques such as the ensemble method [41]. SVM has been used in addition to other algorithms to improve the accuracy of deepfake models [42]. The SVM model was trained using the pre-trained CNN model and achieved an accuracy of 94.85%. Table 3 below shows the SVM performance summary.

Table 3: The SVM Performance Summary

Performance of SVM Model	
Evaluation Metrics	**Performance Results (%)**
Accuracy	94.85
Precision	95.76
Recall	93.72
F1-Score	94.73
Auc Score	98.42

3.3.2.2 Feed-forward Neural Network (FNN)

The FFNN model was trained using the pre-trained CNN Model. FFNN can learn from a pre-trained model [43]. The FFNN model was trained through the transfer learning technique using the pre-trained CNN model and achieved an accuracy of 94.90%. FFNN has powerful nodes that can be trained for depfakes detection tasks [44]. Table 4 below shows the FFNN performance summary.

Table 4: The FFNN Performance Summary

Performance of FFNN Model	
Evaluation Metrics	**Performance Results (%)**
Accuracy	94.90
Precision	94.35
Recall	95.65
F1-Score	95.00
Auc Score	98.62

3.3.2.3 Gated Recurrent Unit (GRU)

The GRU model was trained and evaluated using the pre-trained CNN Model. GRU can capture image features [45]. An accuracy of 94.50% was achieved. Table 5 below shows the performance summary of the GRU model.

Table 5: The GRU Performance Summary

Performance of GRU Model	
Evaluation Metrics	**Performance Results (%)**
Accuracy	94.50
Precision	94.48
Recall	94.66
F1-Score	94.57
Auc Score	98.62

3.4 Ensemble Creation (Ensemble Averaging)

Ensemble averaging within the realm of deep learning is a strategy aimed at enhancing a model's overall performance and resilience by consolidating predictions from several neural networks. Instead of relying on a single neural network, ensemble averaging entails the training and utilization of multiple neural networks that may have different architectures, initializations, or subsets of training data. The predictions generated by these individual networks are subsequently combined or averaged in a particular manner to yield a final prediction. Ensemble averaging stands as a valuable approach for elevating the performance and reliability of deep learning models [46]. The individual model performance accuracies were improved using the Ensemble averaging techniques. The strategies used in the development of the proposed model are discussed in subsequent subsections.

3.4.1 Experimental Method

In this present section, the experimental approach utilized in the building of the social media deepfakes image detection model is presented. In the construction stages, both the pre-processing and feature selection of the data were done using Jupiter Notebook in the Kaggle online development environment. Kaggle is an internet platform that offers a cloud-based programming environment for executing Jupyter Notebooks. It enables users to write, edit, and run Python code in a browser without necessarily needing to install any Python development app on a local computer. Kaggle notebooks are executed directly from their cloud-based server, making it easier for code collaboration, notebook sharing, and working on projects from various devices and locations. The Kaggle platform offers high-speed Processors like the Central Processing Unit (CPU), Graphic Processing Unit (GPU) and Tensor Processing Unit (TPU). The dataset was split into train, validation and test set with a Comma-Separated Values (CSV) file that contains the paths and label of the images. The path column gives the full information about the location of the images and the label gives information on whether the images were real or fake. The train and validation set were used in the training of the model while the test set was used to check the performance of the model. The algorithms used in training the model are CNN, SVM, FFNN and GRU. The model coding, training, and testing were done using Python programming language on the Kaggle online development environment. The local system used was running on Windows 10 x64-bit operating system, 2.1GHZ Intel processor with 4GB RAM Capacity and 500HDD memory capacity. The next subsection discusses the evaluation matrices used.

3.4.2 Model Evaluation

Evaluation matrices are generally used to ascertain the performance and effectiveness of a model in different tasks including classification, regression, and information retrieval. It gives a homogenized procedure to analyze, measure and compare various models [49]. Some of the commonly used evaluation matrices are accuracy, precision, recall, and F1 score [50]. The metrics utilized in the model evaluation include:

3.4.2.1 Precision Matrix

Precision is the ratio of true positives to the sum of true and false positives. Precision is the accuracy of positive predictions. Precision = TP/(TP + FP) [51]. (1)

Where TP = True Positive, TN = True Negative, FN = False Negative, and FP = False Positive.

3.4.2.2 Recall Matrix

Recall is the ability of a classifier to identify all positive instances of every class; it is defined as the ratio of true positives to the sum of true positives and false negatives. Recall = TP/(TP+FN) [52]. (2)

3.4.2.3 F1- score Matrix

F1-score is a weighted harmonic mean of precision and recall such that the best value is 1.0 and the poorest value is 0.0. The weighted average score of F1-score is usually used to compare classifier models, not global accuracy. F1 Score = 2*(Recall * Precision) / (Recall + Precision) [53]. (3)

3.4.2.4 Accuracy Matrix

Accuracy computes the number of correct predictions [50]. All true positive and negative cases are divided by the total number of all cases. The formula is given as (TP + TN)/(TP + FP + FN + TN) [54]. (4)

4. RESULT AND DISCUSSION

This section presents and discusses the outcome of the ensemble technique, a comparative analysis of the individual models and the ensemble, comparative analysis of the ensemble and existing deepfake image models. The ensemble confusion matrix, precision, recall, and accuracy were generated as the main evaluation tool.

4.1 Comparative Analysis of Ensemble Method Result and the individual models

The ensemble technique achieved an accuracy of 94.91%. Table 6 shows the Ensemble performance Summary, Table 7 shows the comparison between the Individual model performance and the ensemble. Figure 4, 5, 6, 7, and 8 shows the confusion matrix of CNN, SVM, FFNN, GRU and Ensemble. The graphical representation of Table 7 is shown in Figure 9.

Table 6: The Ensemble Performance Summary

Performance of SVM Model	
Evaluation Metrics	**Performance Results (%)**
Accuracy	94.91
Precision	95.56

Recall	94.18
F1-Score	94.87
Auc Score	98.80

Table 7: Comparative Analysis of Ensemble Method Result and the individual models

Models	**Evaluation Matrices**				
	Precision (%)	**Recall (%)**	**F1 Score (%)**	**Accuracy (%)**	**AUC Score**
CNN	93.92	95.76	94.83	94.78	98.80
SVM	95.76	93.72	94.73	94.85	98.42
FFNN	94.35	95.65	95.00	94.90	98.62
GRU	94.48	94.66	94.57	94.50	98.62
Ensemble	95.56	94.18	94.87	94.91	98.80

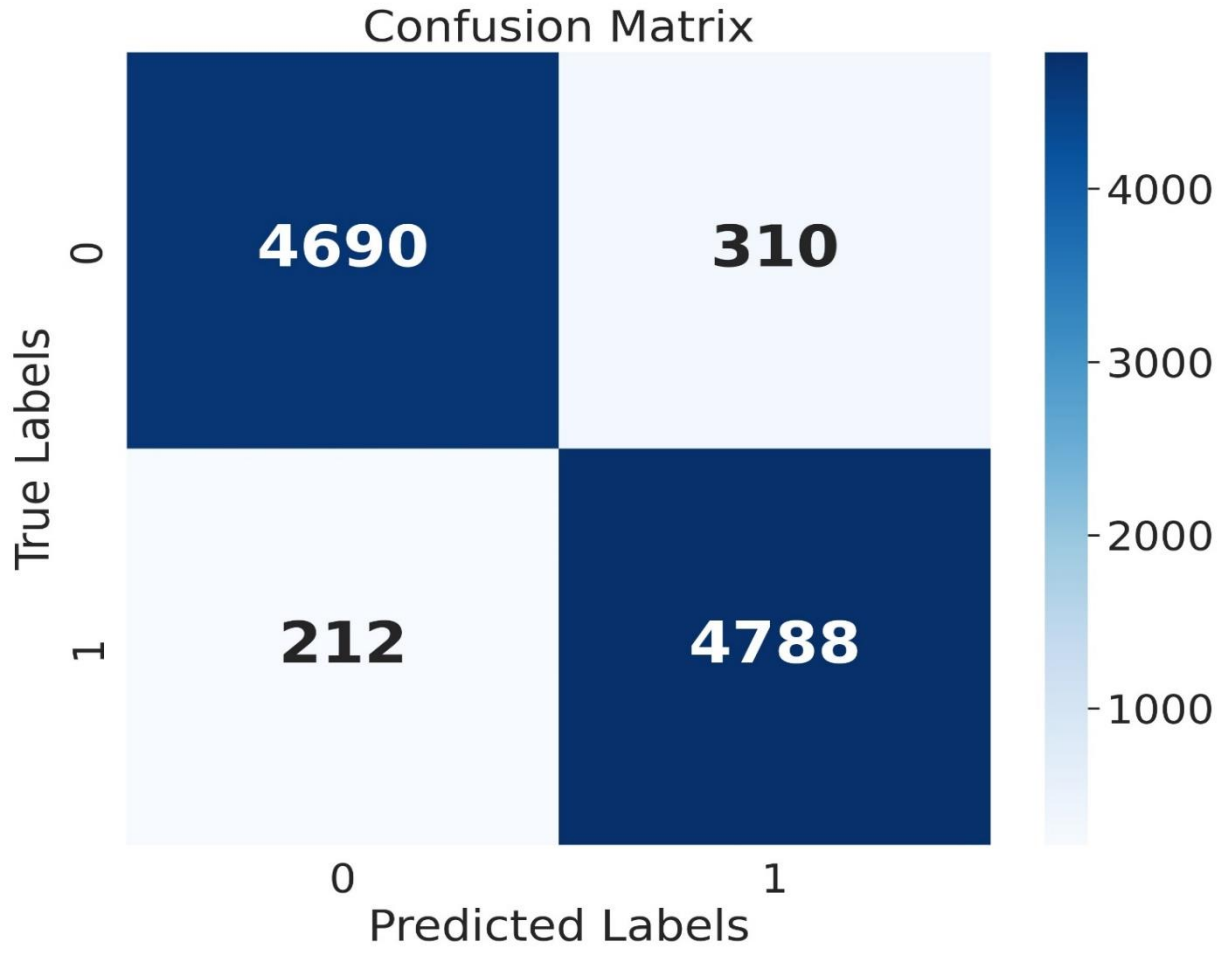

Fig 3: CNN Confusion Matrix

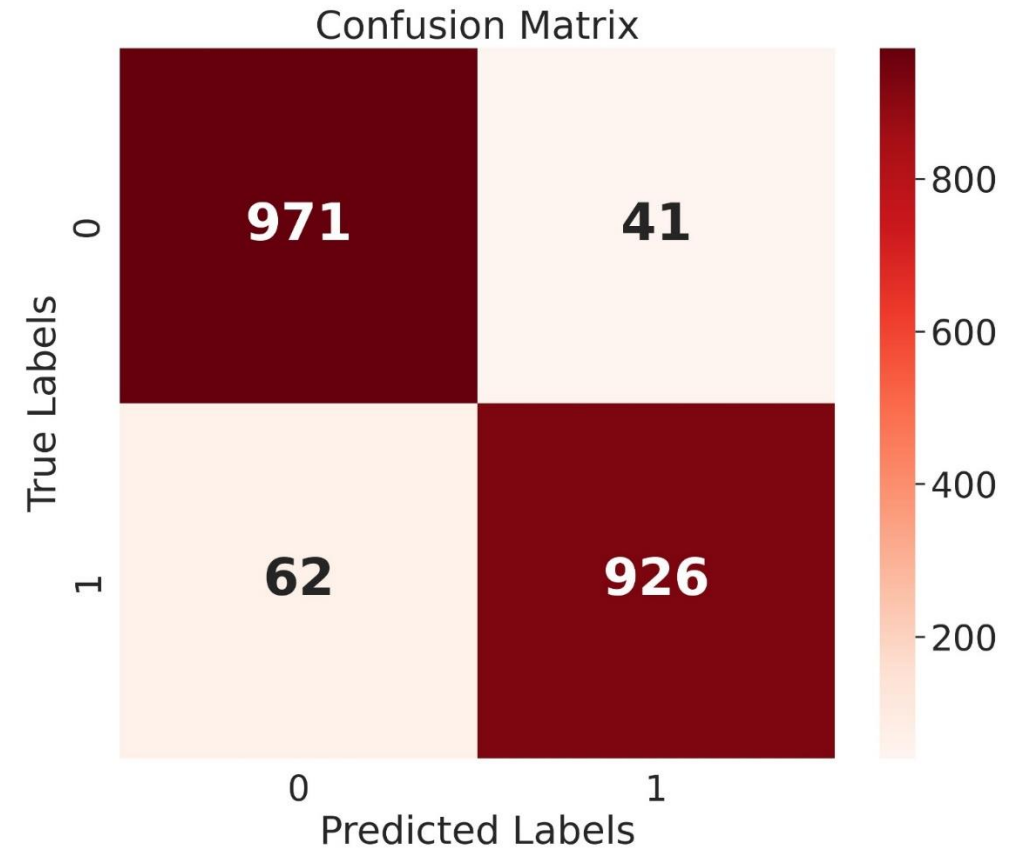

Fig 4: SVM Confusion Matrix

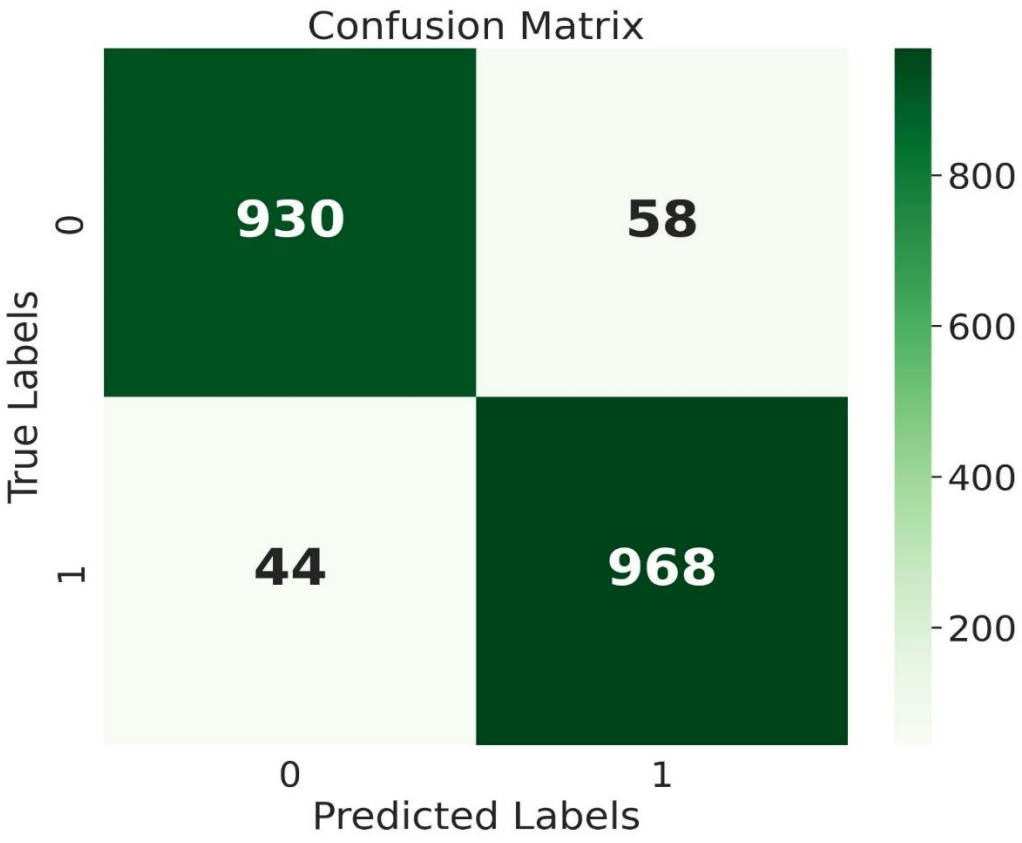

Fig 5: FFNN Confusion Matrix

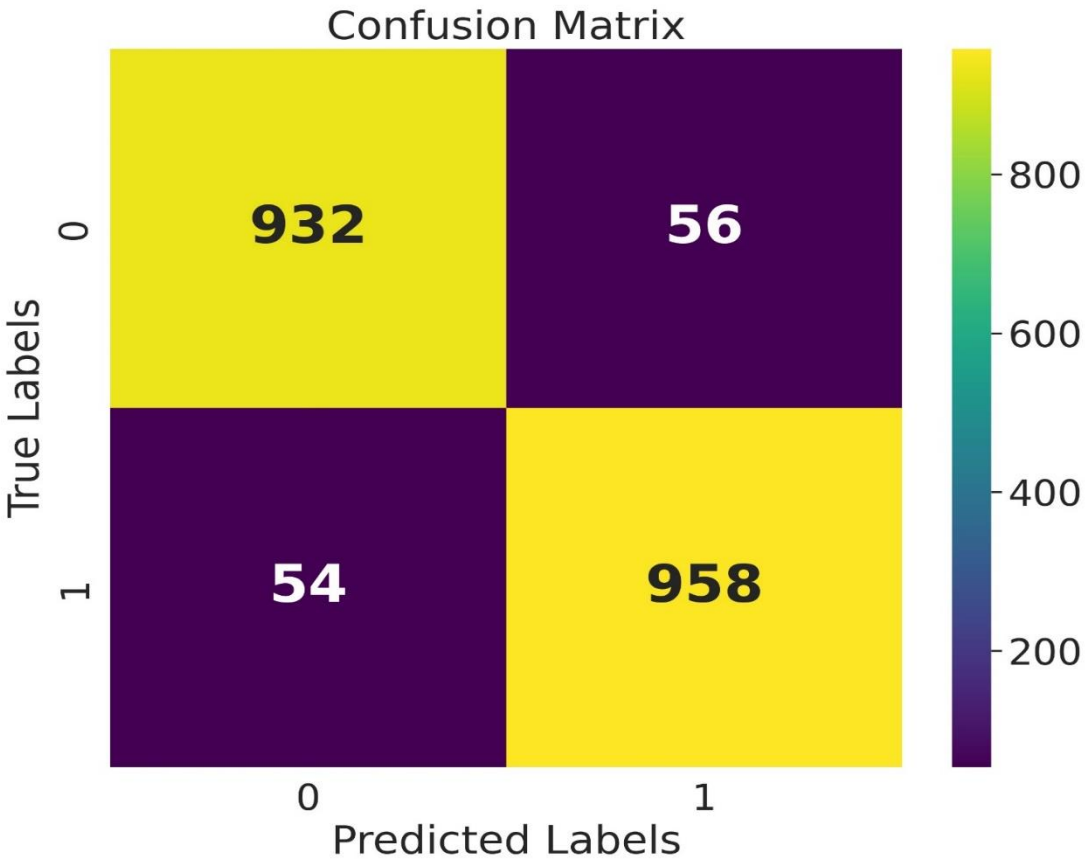

Fig 6: GRU Confusion Matrix

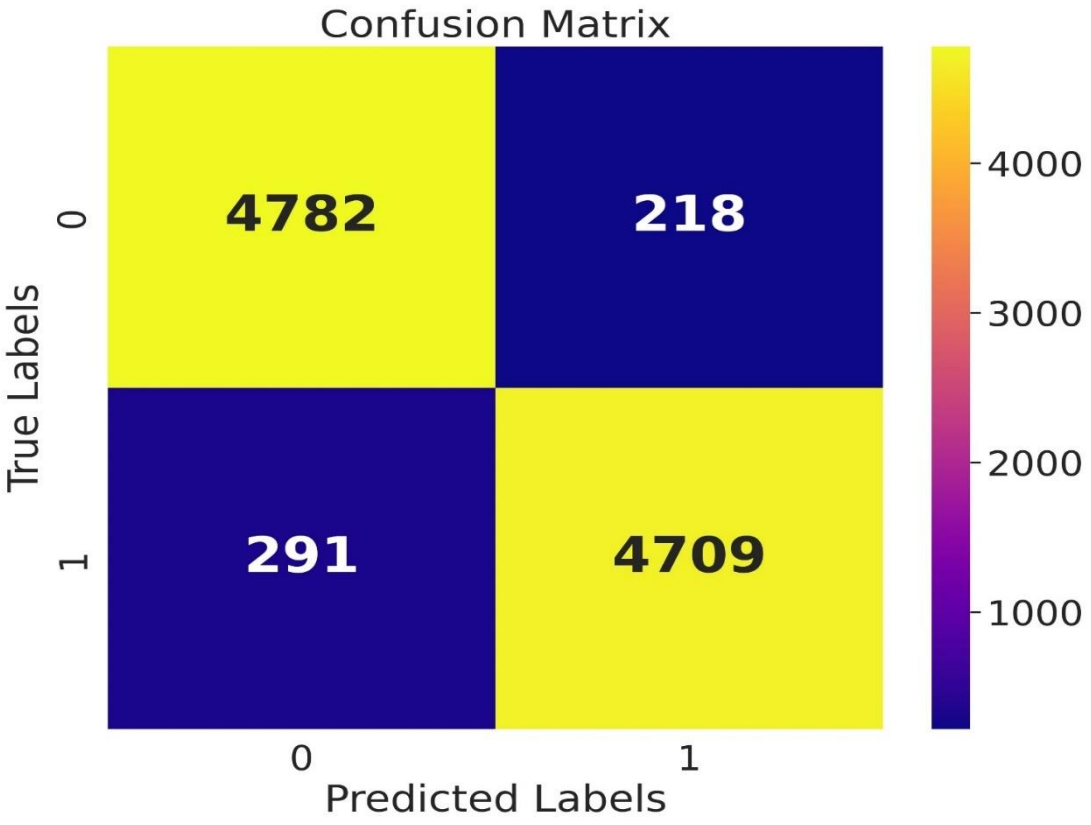

Fig 7: Ensemble Confusion Matrix

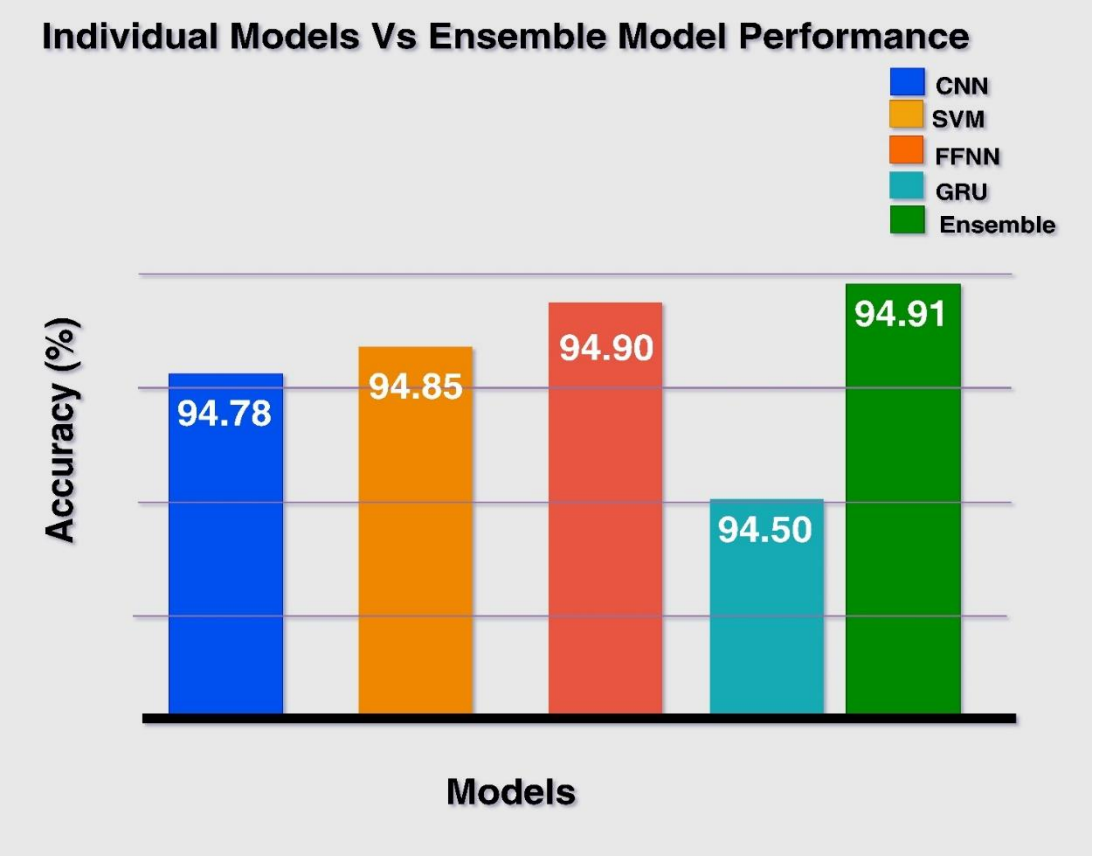

Fig 8: Graphical View of Individual Models and the Ensemble Performance Accuracy

4.2 Comparative Analysis of the Proposed Model and the Existing Models

The performance of the developed model was compared to existing models in image deepfake detection. The research conducted by [27] used CNN and SVM in the development of their model. Moreso, [28] used CNN in the development of their model. Table 8 below shows the summary of the existing model results and the developed model.

Table 8: Comparative Analysis of the developed model with the existing model

Author	Classifier	Accuracy (%)
[27]	CNN	94
	SVM	65
[28]	CNN	94
[25]	DeepFD	94.7
Developed Model	Ensemble	**94.91**

Looking at the table above, the developed model outperforms the existing models in [27], [28] and [25]. The developed model was trained with a higher dataset; the Ensemble technique was utilized to make the model more robust which later achieved a higher performance accuracy. Figure 10 below shows the graphical or pictorial representation of Table 8 above.

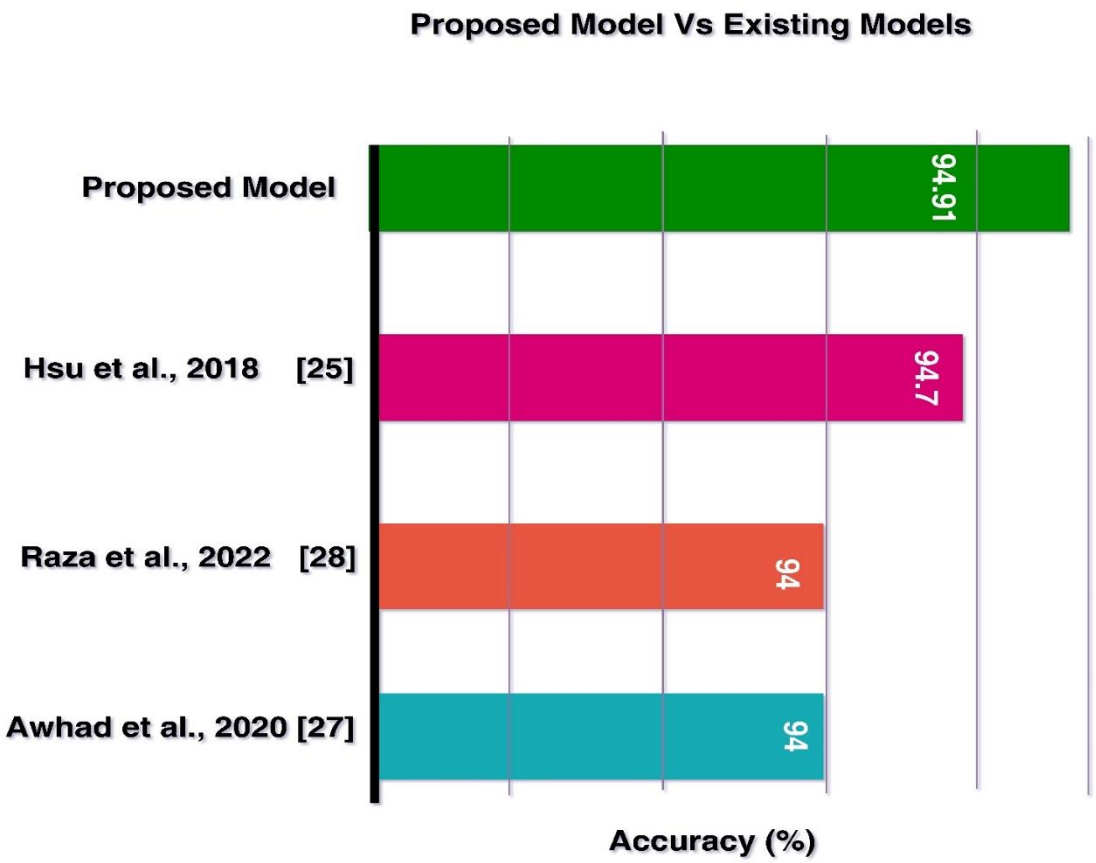

Fig 9: Graphical View of the Proposed Model and Existing Models

5. CONCLUSION

This research work presents a model for the detection of social media image deepfakes. It utilized a publicly accessible dataset on Kaggle. The model was developed using the Kaggle online development environment and Python programming language. Transfer learning technique was utilized to train the SVM, FFNN and GRU from the CNN pre-trained model. The model outperforms the existing model with higher performance accuracy. The model was built using an Ensemble technique making it more robust in detecting deepfake images. The model achieved an accuracy of 94.91%. The dataset employed in the model construction was 70,000 in total, 50% real and 50% fake. The dataset was further divided into train set, validation set and test set. The dataset used is greater than the one used in the existing model [27] and [28]. Due to limited resources, the research is bonded on the resources available in Kaggle's online development environment. The developed model can play an important role in mitigating misinformation on social media especially deepfake image-related misinformation.

6. DECLARATIONS

Acknowledgements

The researchers appreciate all who contributed to the success of this research work.

Conflict of Interest

The authors have no conflicts of interest to declare. The co-authors have seen and agreed with the contents of the manuscript. We certify that the submission is original work and is not under review at any other publication.

Funding:

This study did not receive dedicated funding from any grant-giving organization. The authors conducted this research independently, without any financial backing or support.

Availability of data and material:

The research utilizes the Kaggle dataset, accessible at https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces. This dataset includes both genuine and fake images.

Code Availability:

The code employed for the experimental aspect of the research cannot be publicly shared due to copyright restrictions. Nevertheless, researchers who are interested in accessing the code can do so by sending an email to the corresponding author at (alokejoseph@gmail.com). We are committed to delivering thoughtful responses to inquiries while ensuring adherence to applicable licensing terms and restrictions.

Authors Contribution:

Joshua Abah offered comprehensive guidance and mentorship, while also facilitating research resources, providing valuable insights throughout the study's development, and reviewing the manuscript's final version.

Aloke, Ejike Joseph conceptualized and planned the study, gathered and refined the dataset, conducted dataset analysis, and authored the manuscript.

7. REFERENCES

[1] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "A Novel Machine Learning based Method for Deepfake Video Detection in Social Media," *Proc. - 2020 6th IEEE Int. Symp. Smart Electron. Syst. iSES 2020*, no. May 2021, pp. 91–96, 2020, doi: 10.1109/iSES50453.2020.00031.

[2] A. Boutadjine, F. Harrag, K. Shaalan, and S. Karboua, "A comprehensive study on multimedia DeepFakes," in *2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS)*, 2023, pp. 1–6.

[3] R. K. Nielsen, A. Cornia, and A. Kalogeropoulos, "Challenges and opportunities for news media and journalism in an increasingly digital, mobile, and social media environment," 2016, [Online]. Available: https://rm.coe.int/16806c0385

[4] S. Yadav and N. Tiwari, "Privacy preserving data sharing method for social media platforms," *PLoS One*, vol. 18, no. 1, p. e0280182, 2023.

[5] C. Ziems, Y. Vigfusson, and F. Morstatter, "Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2020, pp. 808–819.

[6] B. Perrigo, "How to Spot an AI Generated Image Like the Balenciaga Pope'," 2023. https://time.com/6266606/how-to-spot-deepfake-pope/ (accessed Jun. 01, 2023).

[7] N. A. Mhiripiri and T. Chari, *Media law, ethics, and policy in the digital age*. IGI Global, 2017.

[8] L. Verdoliva, "Media Forensics and DeepFakes : an overview," pp. 1–24, 2018.

[9] H. Huang, P. S. Yu, and C. Wang, "An Introduction to Image Synthesis with Generative Adversarial Nets," pp. 1–17.

[10] T. T. Nguyen, Q. V. H. Nguyen, C. M. Nguyen, D. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep Learning for Deepfakes Creation and Detection: A Survey," pp. 1–16, 2019, [Online]. Available: http://arxiv.org/abs/1909.11573

[11] B. Paris and J. Donovan, "DEEPFAKES AND CHEAP FAKES," 2019.

[12] C. Bregler, M. Covell, and M. Slaney, "Video Rewrite : Driving Visual Speech with Audio," pp. 1–8, 1997.

[13] J. Vincent, "New AI deepfake app creates nude images of women in seconds," *The Verge*, vol. 27, 2019.

[14] et al I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, "DeepFaceLab is the leading software for creating deepfakes," 2020. https://github.com/iperov/DeepFaceLab (accessed Sep. 09, 2023).

[15] R. Chesney and D. Citron, "Deepfakes and the new disinformation war: The coming age of post-truth geopolitics," *Foreign Aff.*, vol. 98, p. 147, 2019.

[16] R. Brooks, Y. Yuan, Y. Liu, H. Chen, and others, "DeepFake and its Enabling Techniques: A Review," *APSIPA Trans. Signal Inf. Process.*, vol. 11, no. 2, 2022.

[17] A. M. Almars, "Deepfakes detection techniques using deep learning: a survey," *J. Comput. Commun.*, vol. 9, no. 5, pp. 20–35, 2021.

[18] P. Fraga-Lamas and T. M. Fernandez-Carames, "Fake news, disinformation, and deepfakes: Leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality," *IT Prof.*, vol. 22, no. 2, pp. 53–59, 2020.

[19] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. L. Rev.*, vol. 107, p. 1753, 2019.

[20] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv Prepr. arXiv2102.11126*, 2021.

[21] T. Kirchengast, "Deepfakes and image manipulation: criminalisation and control," *Inf. \& Commun. Technol. Law*, vol. 29, no. 3, pp. 308–323, 2020.

[22] U. Scherhag, R. Raghavendra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch, "On the vulnerability of face recognition systems towards morphed face attacks," in *2017 5th international workshop on biometrics and forensics (IWBF)*, 2017, pp. 1–6.

[23] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv Prepr. arXiv1811.00656*, 2018.

[24] E. Meskys, J. Kalpokiene, P. Jurcys, and A. Liaudanskas, "Regulating deep fakes: legal and ethical considerations," *J. Intellect. Prop. Law \& Pract.*, vol. 15, no. 1, pp. 24–31, 2020.

[25] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang, "Learning to detect fake face images in the wild," in *2018 international symposium on computer, consumer and control (IS3C)*, 2018, pp. 388–391.

[26] J. Kim, S. Han, and S. S. Woo, "Classifying genuine face images from disguised face images," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 6248–6250.

[27] R. Awhad, S. Jayswal, A. More, and J. Kundale, "Fraudulent Face Image Detection," in *ITM Web of Conferences*, 2020, p. 3005.

[28] A. Raza, K. Munir, and M. Almutairi, "A Novel Deep Learning Approach for Deepfake Image Detection," *Appl. Sci.*, vol. 12, no. 19, p. 9820, 2022.

[29] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep learning for deepfakes creation and detection," *arXiv Prepr. arXiv1909.11573*, vol. 1, no. 2, p. 2, 2019.

[30] xhlulu, "140k Real and Fake Faces." 2020. [Online]. Available: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces

[31] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, 2020.

[32] R. Rafique, R. Gantassi, R. Amin, J. Frnda, A. Mustapha,

and A. H. Alshehri, "Deep fake detection and classification using error-level analysis and deep learning," *Sci. Rep.*, vol. 13, no. 1, p. 7422, 2023.

[33] A. Ismail, M. Elpeltagy, M. S. Zaki, and K. Eldahshan, "A new deep learning-based methodology for video deepfake detection using XGBoost," *Sensors*, vol. 21, no. 16, p. 5413, 2021.

[34] J. Bernal *et al.*, "Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review," *Artif. Intell. Med.*, vol. 95, pp. 64–81, 2019, doi: https://doi.org/10.1016/j.artmed.2018.08.008.

[35] Y. Patel *et al.*, "An Improved Dense CNN Architecture for Deepfake Image Detection," *IEEE Access*, vol. 11, pp. 22081–22095, 2023.

[36] H. S. Shad *et al.*, "Comparative analysis of deepfake image detection method using convolutional neural network," *Comput. Intell. Neurosci.*, vol. 2021, 2021.

[37] A. Karandikar, V. Deshpande, S. Singh, S. Nagbhidkar, and S. Agrawal, "Deepfake video detection using convolutional neural network," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, pp. 1311–1315, 2020.

[38] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[39] C. Desai, "Image classification using transfer learning and deep learning," *Int. J. Eng. Comput. Sci.*, vol. 10, no. 9, pp. 25394–25398, 2021.

[40] H. Agarwal, A. Singh, and D. Rajeswari, "Deepfake Detection Using SVM," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2021, pp. 1245–1249.

[41] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "A Machine Learning Based Approach for Deepfake Detection in Social Media Through Key Video Frame Extraction," *SN Comput. Sci.*, vol. 2, no. 2, pp. 1–18, 2021, doi: 10.1007/s42979-021-00495-x.

[42] L. A. Passos, D. Jodas, K. A. P. da Costa, L. A. S. Júnior, D. Colombo, and J. P. Papa, "A review of deep learning-based approaches for deepfake content detection," *arXiv Prepr. arXiv2202.06095*, 2022.

[43] J. L. Gayathri, B. Abraham, M. S. Sujarani, and M. S. Nair, "A computer-aided diagnosis system for the classification of COVID-19 and non-COVID-19 pneumonia on chest X-ray images by integrating CNN with sparse autoencoder and feed forward neural network," *Comput. Biol. Med.*, vol. 141, p. 105134, 2022.

[44] P. Le-Hong and A.-C. Le, "A comparative study of neural network models for sentence classification," in *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, 2018, pp. 360–365.

[45] Z. Chang, X. Zhang, S. Wang, S. Ma, and W. Gao, "Iprnn: An Information-Preserving Model For Video Prediction Using Spatiotemporal Grus," 2021, pp. 2703–2707. doi: 10.1109/ICIP42928.2021.9506391.

[46] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, p. 100203, 2019.

[47] N. Altman and M. Krzywinski, "Ensemble methods: bagging and random forests," *Nat. Methods*, vol. 14, no. 10, pp. 933–935, 2017.

[48] H. Drucker, C. Cortes, L. D. Jackel, Y. LeCun, and V. Vapnik, "Boosting and other ensemble methods," *Neural Comput.*, vol. 6, no. 6, pp. 1289–1301, 1994.

[49] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. \& Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[50] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv Prepr. arXiv2010.16061*, 2020.

[51] S. H. Kok, A. Azween, and N. Z. Jhanjhi, "Evaluation metric for crypto-ransomware detection using machine learning," *J. Inf. Secur. Appl.*, vol. 55, p. 102646, 2020.

[52] I. Idrissi, M. Azizi, and O. Moussaoui, "Accelerating the update of a DL-based IDS for IoT using deep transfer learning," *Indones. J. Electr. Eng. Comput. Sci*, vol. 23, no. 2, pp. 1059–1067, 2021.

[53] F. Muharemi, D. Logof\uatu, and F. Leon, "Machine learning approaches for anomaly detection of water quality on a real-world data set," *J. Inf. Telecommun.*, vol. 3, no. 3, pp. 294–307, 2019.

[54] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. data Min. \& Knowl. Manag. Process*, vol. 5, no. 2, p. 1, 2015.