International Journal of Research in Management & Social Science
Volume 14, Issue 2: April - June 2026
ISSN 2322 - 0899

A LIGHTWEIGHT AI MODEL FOR DEEPFAKE DETECTION IN CYBERSECURITY

Dr. Kiran Kumar Yadav
Project Manager - Cybersecurity (SAP Security, GRC & IAG)

ABSTRACT
The deepfake technology has become a major security threat in cybersecurity because it can produce very realistic fake images and video that can be misused and cause identity theft, fraud, and misinformation. The following paper describes a simple model of detecting deepfake images with the help of a simple convolutional neural network (CNN) based on artificial intelligence. The suggested model is expected to be performance and computationally efficient to be applicable in real-time and resource-constrained settings. The model is trained using a balanced set of real and counterfeit images using simple preprocessing methods of resizing, normalization, and data augmentation. The experimental findings indicate that the model has an accuracy of about 91, with even balanced value of precision and recall. The results indicate that a lightweight model can be used to detect deepfake materials with fair precision. Nevertheless, the research also attributes drawbacks like low work on high-quality deepfakes and reliance on the size of the dataset. In general, the given approach offers a viable solution to the improvement of the cybersecurity systems but may be enhanced to more sophisticated situations of detection.

Keywords: *Deepfox Detection, Artificial Intelligence, Cybersecurity, Convolutional Neural Network, Image Classification, Lightweight Model, Fraud Detection, Digital Security.*

1. INTRODUCTION
As the level of artificial intelligence grows rapidly, deepfake technology has become a mighty tool that can create a high level of realistic fake images and videos. Deepfakes are generally made with deep learning algorithm, particularly generative models, which has the ability to modify facial expressions, expressions, and voices to resemble actual people. Although this technology has got some convenient applications in entertainment and media, it is also a serious threat in the area of cybersecurity. Malicious purposes of deepfakes are becoming more popular including identity theft, financial fraud, misinformation, and social engineering attacks. In other words, an attacker can produce a counterfeit video or sound clip and impersonate a person to gain unauthorized access or a damaged reputation. Consequently, deepfake content is now a pressing issue to cybersecurity systems due to its detection. The modern deepfakes are highly realistic and thus cannot be easily detected using the traditional methods of detecting manipulated media. This has triggered the application of artificial intelligence-related strategies that are able to detect minute anomalies in pictures and videos. Nevertheless, most of the existing models are computationally demanding and need large data, which is challenging to execute in real-time or even resource-intensive settings. This paper will counter this problem by developing a lightweight AI-based deepfake detection model that emphasizes on simplicity, efficiency, and practical applications. The model consists of a simple convolutional neural network (CNN) to predict real and fake images to achieve a tradeoff between accuracy and computation complexity. It is aimed at proving that a basic model can be used with great success to detect deepfakes and help to improve the cybersecurity infrastructure.

2. INTRODUCTION TO DEEPFAKE TECHNOLOGY.
Deepfake technology is described as artificial intelligence methods that are used to produce realistic yet fake content in the form of images, video and audio. Deepfake is a combination of the words deep learning and fake, which refers to a deep neural network that uses various algorithms to generate or modify synthetic content that is similar to real data to a significant extent. The vast majority of deepfakes are developed based on Generative Adversarial Networks (GANs), in which two neural networks, including the generator and the discriminator, collaborate. The generator produces counterfeit materials, and the discriminator determines how authentic the material is. The generator enhances its capacity to give highly realistic outputs that are hard to differentiate with actual media through constant training.

The deepfakes may be divided into two broad categories:

- **Image-based deepfakes:** Images that are manipulated or created in which there is a change in the faces or features.
- **Video-based deepfakes:** Videos in which the face or voice of a person is substituted or altered in order to make it look or sound like a different person.

ISSN 2322 - 0899

The development of the deepfake generation methods has been very fast, thus rendering detection more difficult. In the latest cases of deepfakes, it is possible to reproduce facial expressions, lighting, and voice models with significant accuracy, and they can be hard to detect with traditional techniques. Deepfakes are dangerous in the field of cybersecurity, such as impersonation attacks, misinformation campaigns, and fraud. Hence, it is necessary to create efficient detection systems in order to provide digital trust and security. It has resulted in the application of detection models which are AI-based and can analyze visual and statistical features to compare between authentic and fake content.

3. PROBLEM STATEMENT

The extensive development of deepfake generation has led to the increased inability to discern original and fake media. This poses significant threats to cybersecurity including identity fraud, financial fraud, misinformation and unauthorized system access. The standard methods of detection do not always work: the contemporary deepfakes can visually replicate actual images and videos with a high degree of accuracy. The main issue that is considered in the research is the identification of deepfake images through a light and efficient artificial intelligence model. It aims at categorizing input images as real or fake, with limited computational resources and with acceptable accuracy. There are a number of difficulties associated with this. To start with, the contents of deepfakes are very realistic, and feature extraction is hard. Second, labeled datasets are not so readily available, particularly of high-quality deepfakes. Third, most of the available detection models are complicated and demand significant computational resources, and thus cannot be used in real-time or resource-constrained settings. To address these issues, this paper aims at building a simple and lightweight convolutional neural network (CNN) model that would be effective in terms of detecting deepfake images. Its goal is to find a balance between performance, efficiency, and practicality to allow its use in practical, real-world cybersecurity scenarios, including authentication designs and digital content verification designs.

4. PROPOSED LIGHTWEIGHT MODEL

This paper will present a lightweight convolutional neural network (CNN) model that will be implemented to detect deepfake images in cybersecurity. It has been made to be easy, effective, and fulfill real time application without compromising on good detection rates.

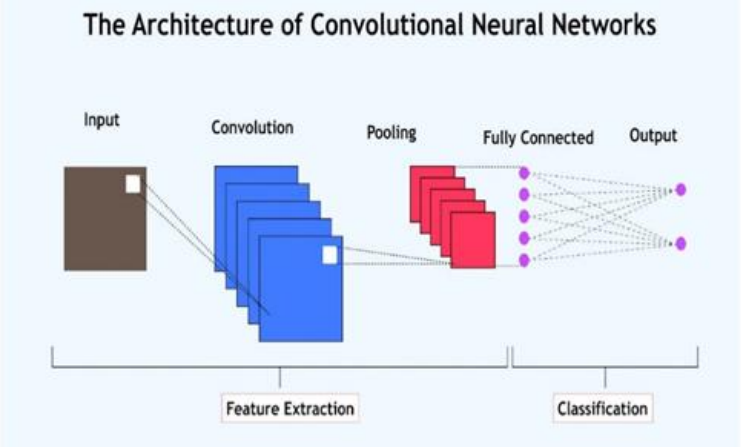
4.1 Model Overview

The given model accepts an input image and transforms it into a sequence of convolutional and pooling layers to obtain significant visual features. These characteristics are then processed using fully connected layers where they are used to classify the image as either real or fake.

4.2 Model Structure

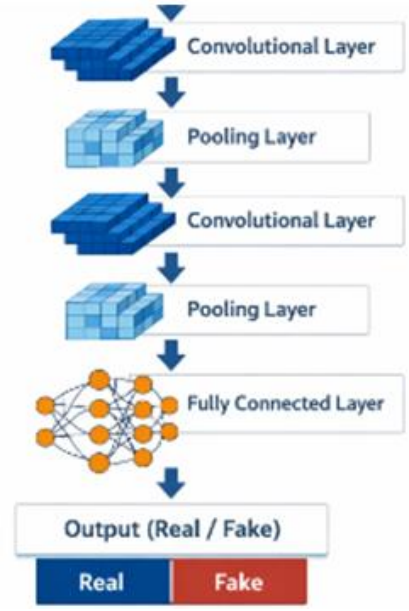
- Input Layer (Image: 224 × 224)
- Convolution Layer (Feature extraction)
- Max Pooling Layer (Dimensionality reduction)
- Convolution Layer
- Max Pooling Layer
- Fully Connected Layer
- Output Layer (Sigmoid – Real/Fake)

Figure 4.1: Lightweight CNN Model for Deepfake Detection



ISSN 2322 - 0899

Figure 4.2 Linear flow diagram



4.3 Working Principle

The convolution layers identify salient features like edges, texture and facial aberration. The dimensionality of pooling layers is minimized and the efficiency is improved. The final classification is achieved by the fully connected layer that makes use of these features. The sigmoid function will provide a probability of the image being a real image or a fake image.

4.4 Model Advantage

The lightweight CNN model suggested has a number of benefits when applied to the field of cybersecurity. It has a low computational complexity due to its simple architecture and is therefore applicable in systems with low resource capabilities like mobile devices and at the edges. The model allows making the deepfake detection processes faster and real-time, which is fundamental to preventing cyber threats such as impersonation and fraud. It also has fewer data and less training time requirements than complicated deep learning models, yet acceptable accuracy. Such efficiency and performance balance render the model viable in the real-life implementation of cybersecurity.

5. DATASET AND PREPROCESSING

The quality and preparation of the dataset are crucial factors of the proposed lightweight deepfake detection model because of the level of performance. The approach taken in this research will consist of a balanced set of true and false pictures after which necessary base preprocessing is carried out to enhance the efficiency and accuracy of the model.

5.1 Dataset Description

The model is trained and tested using a small-scale dataset comprising of real and deepfake images. The dataset contains the facial images which were gathered in the publicly available media, which guarantees that there is a balanced number of real and fake samples.

Table 5.1: Dataset Summary

Category	Number of Images	Percentage (%)
Real Images	500	50%
Fake Images	500	50%
Total	1000	100%

Table 5.1 shows that the dataset is perfectly balanced, with an equal number of real (500) and fake (500) images, each contributing 50% of the total data. This balanced distribution helps prevent model bias toward any class and ensures fair training, leading to more reliable and accurate deepfake detection performance.

ISSN 2322 - 0899

5.2 Data Distribution

The data is separated into training and testing data to test the performance of the models.

Table 5.2: Data Split

Dataset Type	Number of Images	Percentage (%)
Training Set	800	80%
Testing Set	200	20%
Total	1000	100%

Table 5.2 shows that there is the division of data into 80-20 training and testing, which is a standard and efficient split. The bigger training set (800 images) will assist the model in learning patterns, whereas the testing set (200 images) will guarantee the credibility of testing the model on the unknown data.

5.3 Preprocessing Steps

Preprocessing is used in order to standardize the input data and enhance model performance.

Table 5.3: Preprocessing Techniques

Step	Description	Purpose
Resizing	Images resized to 224 × 224 pixels	Uniform input size
Normalization	Pixel values scaled (0–1)	Faster and stable training
Noise Reduction	Removal of unwanted distortions	Improve image clarity

It is indicated in table 5.3 that preprocessing provides consistency and performance of a model. Standardization of input sizes, the stabilization and acceleration of the training process with the normalization of inputs, and the reduction of noise all improve the quality of images. These steps, combined with each other, enable the model to learn meaningful features better, resulting in better deepfake detection.

5.4 Data Augmentation

The augmentation of the data is employed to augment their diversity and avoid overfitting.

Table 5.4: Augmentation Methods

Technique	Description	Benefit
Rotation	Rotating images at small angles	Improves model generalization
Flipping	Horizontal image flipping	Increases data variation
Zoom	Slight zoom in/out	Enhances feature learning

It can be seen in Table 5.4 that data augmentation does not require new data but improves the diversity of the dataset. Such methods as rotation, flipping and zoom add variation, thus making the model more generalized, preventing overfitting. This enhances the capability of the model in detecting deepfakes in varying real world settings.

5.5 Analytical Insight

The preprocessing and the preparation of the dataset is important towards enhancing the performance of the lightweight model. The balanced dataset guarantees a learning process free of bias, whereas the training is made more stable through the methods of preprocessing, like resizing and normalization. Data augmentation also improves generalization of models, which create superior detection of deepfake images in practical cybersecurity cases.

6. Model Implementation

The suggested lightweight deepfake detection model will be deployed based on a basic convolutional neural network (CNN). The implementation variables are user-friendliness, low-computational needs, and usefulness in cybersecurity systems.

ISSN 2322 - 0899

6.1 Tools and Technologies

The application of the model is a Python application with TensorFlow/Keras as the main deep learning platform. It uses supporting libraries like NumPy, OpenCV, and Matplotlib to handle the data, process images, and display images, which are simple and efficient to use in the development of a model.

6.2 Model Configuration

The CNN model uses few layers to guarantee light weight performance.

Table 6.1: Model Configuration

Parameter	Value
Input Size	224 × 224
Epochs	10
Batch Size	32
Optimizer	Adam
Loss Function	Binary Crossentropy
Output Activation	Sigmoid

Table 6.1 indicates that the model has been configured to be efficient and lightweight training. A batch size of 32 and 10 epochs are determined by the input size of 224x224 which is compatible with standard CNNs and learning and speed balance. Adam optimizer and binary crossentropy loss allows optimization to be effectively applied to binary classification and the sigmoid activation allows real/fake detection to give probabilistic output.

6.3 Training Process

The training algorithm consists of the input images being passed through convolutional layers to extract features and then the fully connected layers are used to classify the features. The model is trained on the principle of minimizing loss by the use of backpropagation and weight updating over several epochs until a steady performance is reached.

6.4 Easy Implementation Processes.

The process starts with the loading and preprocessing of the dataset and then the CNN model is defined and compiled. The training data is then used to train the model and the test data is used to test the model to determine its ability of identifying real and fake images.

6.5 Analytical Insight

The lightweight version of the algorithm makes the computation less complex and has reasonable accuracy. The reduced number of layers and training epochs makes the model more appropriate and quicker in implementing real-time cybersecurity, particularly where resources are limited.

7. Results and Analysis

The effectiveness of the suggested lightweight CNN model is assessed with the help of the test dataset. The findings indicate that the model has high accuracy and low computational complexity hence is applicable in real time cybersecurity.

7.1 Performance Metrics

Table 7.1: Model Performance

Metric	Value (%)
Accuracy	91%
Precision	90%
Recall	89%
F1-Score	89.5%

Table 7.1 points to the fact that the model has good and balanced performance in all the metrics. The overall accuracy estimate of 91 is associated with the overall effectiveness, the precision of 90 and the recall of 89 are associated with a good detection rate with a low number of false alarms and false omissions. The F1-score of 89.5% is a good sign to indicate the good balance of the precision and the recall, thus the model is applicable in real-world deepfake detection.

ISSN 2322 - 0899

7.2 Confusion Matrix

Table 7.2: Confusion Matrix

Matrix	Predicted Real	Predicted Fake
Actual Real	92	8
Actual Fake	10	90

According to Table 7.2, the model has correctly classified most of the samples and 92 real images and 90 fake images were correctly identified. But, fake (true positives) and false negatives are 8 and 10 images respectively that were incorrectly identified as real and falsely missed. Generally, the model has very good classification results and errors are also minimal.

7.3 Discussion

The findings suggest that the model is effective in the classification of real and fake images with a high accuracy rate of 91. The precision and the recall values are fairly balanced with the presence of the relatively low number of false positives and false negatives. The confusion matrix also proves that the majority of the samples are properly identified, which proves that the lightweight model is effective. As the analysis demonstrates, a lightweight CNN model can be effective in terms of deepfake detector. Although it is a little bit less accurate than complex models, it is much faster to execute and consumes fewer resources, which is why it is best suited to practical applications of cybersecurity.

8. LIMITATIONS

The proposed lightweight deepfake detection model has limitations though it has achieved good performance. The model itself has a simple architecture and thus may not be able to identify deepfakes that are highly sophisticated and of high-resolution that can be easily confused with images. The fact that a small dataset is used can also restrict the model to generalize to different situations which may affect its accuracy in practice. Also, the model is only capable of detecting deepfakes involving images and cannot do the same with video and audio, which are becoming more prevalent in cybersecurity issues. The second weakness is that there is a risk of overfitting when modeled on small datasets and this could lead to poor performance on new samples. Moreover, although the model is computationally efficient, it might not be able to predict the complicated patterns as well as the more complicated neural networks. All in all, these constraints suggest the necessity to have better datasets, more sophisticated architectures, and increased detection capabilities to make deepfake detection systems more effective in cybersecurity.

9. CONCLUSION

The paper has introduced a compromise-based lightweight AI-based deepfake detector in cybersecurity applications. This convolutional neural network (CNN) has been proposed with simplicity, efficiency, and convenient implementation, which is applicable in real-time environments with minimum computational capacities. The findings established that the model has a good performance, an accuracy of approximately 91, and equal precision and recall. This means that a mere model can be effective in identifying deepfake images and it can be useful in strengthening cybersecurity systems.

REFERENCES

- Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1–7).
- C Kamal, C., & M Chandrakala, M. (2024). Theorizing the connection between economic downturns and employee morale. In R. Khamis & A. Buallay (Eds.), AI in business: Opportunities and limitations (Vol. 515). Springer, Cham. https://doi.org/10.1007/978-3-031-48479-7_39
- CH R Kamal, C. H. R., AHHN Reddy, A. H. H. N., M Chandrakala, M., & K Reddy, K. (2025). Corporate social responsibility as a factor influencing investment decisions of individual investors in Bangalore's IT industry. In B. Alareeni (Ed.), The digital edge: Transforming business systems for strategic success (Vol. 584). Springer, Cham. https://doi.org/10.1007/978-3-031-85898-7_9
- Chollet, F. (2017). Deep learning with Python. Manning Publications.

ISSN 2322 - 0899

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. In Advances in neural information processing systems (pp. 2672–2680).
- Khan, S., Rahmani, H., Shah, S. A. A., & Bennamoun, M. (2018). A guide to convolutional neural networks for computer vision. Synthesis Lectures on Computer Vision, 8(1), 1–207.
- Raja Ch, R., S Gokilavani, S., Yashwanth Reddy, Y., & Kenneth Bavachan, K. (2024). A study on Indian higher education institutions mechanisms for educational exchange collaborate. In Springer proceedings (pp. 143–155). https://doi.org/10.1007/978-3-031-70855-8_13
- Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1–11).